8263-EN-01
DTIC
Prof S. Openshaw
U of Leeds UK

Proceedings of

the second annual conference

GeoComputation 97

Incorporating the
Spatial Information Research Centre's
9th Annual Colloquium

26-29 August 1997

University of Otago, Dunedin, New Zealand

OF LEEDS

SPATIAL
INFORMATION
RESEARCH
CENTRE

SAPERE·AUDE

Edited by Richard T. Pascoe

Best
Available
Copy

Proceedings of

the second annual conference

GeoComputation 97

Incorporating the
Spatial Information Research Centre's
9th Annual Colloquium

26-29 August 1997

University of Otago, Dunedin, New Zealand

OF LEEDS

SPATIAL INFORMATION RESEARCH CENTRE

SAPERE · AUDE

Edited by Richard T. Pascoe

AIR NEW ZEALAND Price Waterhouse Urwick

MapInfo

Financial support for this conference has been provided from the European Research Office of the U.S. Army.

Members of the organising committees would like to express their sincere appreciation for the support offered by the sponsors listed above.

FRONT COVER: A visualisation to support tasks in minerals exploration, showing a combination of different geophysical data sources.

The picture represents a single image captured from a fully interactive environment built using IRIS Explorer, and was produced as part of the geographic data visualisation project, at Geographic Information Science, Curtin University of Technology, Australia.

This and further examples are available from http://www.cs.curtin.edu.au/gis/visualisation/)"

Additional copies of these proceedings may be obtained from:
Dr. Richard Pascoe
Manager
Spatial Information Research Centre
University of Otago
PO Box 56
Dunedin
New Zealand

Email: rpascoe@commerce.otago.ac.nz
Phone: +64 3 479 8321
Fax: +64 3 479 8311

August 1997

ISBN 0-473-04564-8

Computing and the science of Geography: the postmodern turn and the geocomputational twist

W.D.Macmillan
School of Geography,
University of Oxford, Mansfield Road,
Oxford OX1 3TB, U.K.
email: bill.macmillan@geog.ox.ac.uk

AIR NEW ZEALAND

Air New Zealand Guest Keynote Speaker

Presented at the second annual conference of GeoComputation '97 & SIRC '97, University of Otago, New Zealand, 26-29 August 1997

Introduction

In the wake of Kuhn's (1962) attack on established notions of scientific progress, Haggett and Chorley (1967) announced that Geography was undergoing a quantitative revolution. In fact, it was more of a battle cry than an announcement. It stirred up rebellion within the discipline and sent marauders off into neighbouring domains to bring back intellectual booty. Like the quantitative revolution, geocomputation is an enterprise stretching well beyond the borders of academic geography. The two movements have many other characteristics in common but they also have import differences, the most significant of which is the most obvious - a radical difference in accessible computing power.

In the period between the heyday of the quantitative revolution and the coining of the term geocomputation, much philosophical water has flowed under the geographical bridge. There have been major surges from humanism, Marxism and, latterly, postmodernism and there have been many minor currents. But throughout this period, the geocomputational tide has been rising, little noticed by the philosophers of geography. Much of their concern, as proponents or opponents, has come to focus on the 'postmodern turn'. Until recently, they have largely ignored the geocomputational twist in the tail of quantitative geography - or in what they had taken to be its tail.

This paper seeks to place the 'geocomputational twist' in its philosophical and historical setting, stimulated in part by a series of email exchanges between the organisers of GeoComputation '97 on possible definitions of the neologism. It presents an illustrated argument in support of two propositions: that the quantitative revolution and the burgeoning of computational geography belong to the same, long-standing, intellectual tradition; and that that tradition [illegible]ing. One could argue, as hinted above, that [illegible]putation is a continuation or addendum to the quantitative revolution but one can equally well view the latter as a rehearsal for the former. If one takes this position - standing, as it were, at the present looking back - then it is clear that the rehearsals were under way well before the 1960s. It is equally clear that geocomputation, when looked at in these terms, has a long way to go before it fulfils its promise.

The paper sketches out a few ideas on the foundations of scientific geography, where the latter term is taken to embrace rather more than the academic discipline. It looks briefly at measurement, calculation and computing technology prior to the electronic age, using Harrison's chronometers, the Varignon Frame, and the notion of market equilibrium as examples. It presents a thumb nail sketch of the standard picture of science in the quantitative revolution and of the social context within which scientific geography was promoted. And it considers certain counter-revolutionary criticisms of the notion that society may be studied scientifically.

An important part of the argument is a consideration of the changes that have occurred in science since the standard, physics-based, picture was painted. That picture was always a caricature. The expansion of the biosciences, the explosion of interest in nonlinear systems in general and chaos in particular, the associated discovery of the fundamental unpredictability of certain physical and biological systems, and the recognition that objectivity in science is a direction rather than a terminus have all contributed to the blurring of the supposed science/social science distinction. And at the centre of much of this change has been computing. It was, after all, in the humming of a Royal McBee that Edward Lorenz first detected chaotic behaviour.

Such behaviour might be thought to be a recent characteristic of the discipline itself, or, perhaps, of its philosophically self-conscious branches. But under the postmodern froth there is a strong geocomputational brew. Emblematically, whilst the revisionary metaphysicians have been exercised about the notion of truth, the spatial scientists have been harnessing fuzzy logic.

As for the social context, it has, of course, changed radically since the '60s. And those changes, as any good materialist should admit, have all but put paid to the Marxist project. Such force as there was in Harvey's (1989) accusation that modellers have produced 'little more than the proverbial hill of beans' has been eclipsed by the collapse of the house of cards that represented the Marxist project in practice. Of at least equal significance, arguably, has been the extraordinary advancement in computing power, the emancipatory effect of its widespread availability, and the wiring of society.

Drawing these threads together, the paper attempts both to justify the claims made about the methodological significance of the geocomputational twist and to highlight the shortcomings in the contemporary portfolio of geocomputational activities.

The Analytic Tradition

One of the difficulties inherent in understanding the debate about the nature of the quantitative revolution – and, by extension, the nature of geocomputation - is a persistent and often wilful misuse of terms. The words 'quantitative' and 'revolution' both require scrutiny, as does the term 'positivism'. As Taylor and Johnston (1995 p.52) have argued, the quantitative revolutionaries adopted markedly different approaches and had different views on an appropriate name for their movement:

> three early popular labels were "conceptual", "model-based", and "statistical" – before the label "quantitative" was generally adopted

This heterogeneity has been played down by critical historians who have found it convenient to use a single label and to ascribe a particular view of science to those it has been attached to (*ibid.* p.52):

> The quantifiers were criticised from a range of contrary positions for their excessively narrow interpretation of what constitutes science. In this process the quantitative revolution was reconstructed as a unitary monolith (sic) and any diversity associated with its theoreticians tended to be written out of the story.

Taylor and Johnston go on to argue that there were tensions within the movement (*ibid.* p.52)

> between deductive and inductive "science" and ... between "pure" and "applied" geography

And they say that in the early stages (up to the 1970s) it was pure geography that dominated. Thus, at least in the first flush of the quantitative revolution, geography had some resemblance to the standard model of a science, with rationalist and empiricist wings and what Taylor and Johnston refer to as a 'mainstream concern for models and theory'. Arguably, then, 'scientific' would be a better label than 'quantitative'.

The term 'revolution' is not particularly illuminating either. As the introduction to this paper suggests, its early use in geography was as much prescriptive as descriptive. The extent to which the movement actually was revolutionary is a matter for debate, as is Kuhn's view about the nature of change in science. What is more, there appears to be a mismatch between Kuhn's conception of science and the views of the revolutionary geographers about their own

work. They seem to have subscribed to the idea that science is a rational and cumulative enterprise, which deals objectively with testable propositions about the real world. Kuhn challenged this idea. As Searle (1996, pp.11-12) puts it:

> Kuhn sometimes seems to be arguing that there is not any such thing as the real world existing independently of our scientific theories, which it is the aim of our theories to represent. Kuhn, in short, seems to be denying realism.

He then adds (*ibid.* p.12):

> Most philosophers do not take this denial of realism at all seriously. Even if Kuhn were right about the structure of scientific revolutions, this in no way shows that there is no independent reality that science is investigating.

Whilst the quantitative revolutionaries were happy to appeal to Kuhn's ideas to justify their attempts to transform the discipline, few if any shared his relativism. Behind the rhetoric of scientific revolution - derived from arguments about revolutionary change *within* a science - was a more gradual but in some ways more profound transition from an unscientific to a partially scientific geography.

As for 'positivism', it is seldom clear what various users of the term have in mind, apart for their disapproval. In the philosophical literature, 'positivism' tends to be used, if at all, as a contraction of 'logical positivism'. The nature of this school is neatly summarise by Solomon (1997, p.720):

> The main thrust of logical positivism is its total rejection of metaphysics in favour of a strong emphasis on science and verifiability through experience. The method of the logical positivists, accordingly, is strongly empiricist (*they actually called themselves "logical empiricists"*)...

In addition to the rejection of metaphysics, the logical positivists had a clear view about ethical and aesthetic statements. They thought (Pettit, 1993, p.9) that:

> Evaluative propositions did not serve, or at least did *not serve primarily, to essay a belief as to how things* are; their main job was to express emotion or approval/disapproval, much in the manner of an exclamation like 'Wow!' or 'Ugh!'

The logical positivists, then, were concerned with 'how things are' and they took the view that evaluative statements do not help *with this task*. But there is a great deal of sloppy reasoning between that observation and the notion that 'positivists', in some ill-defined sense, are not concerned with matters of conscience or social justice. And the reasoning is worse than sloppy when it comes to suggesting, as some recent *geographical writing appears to* do, that positivists, *qua* positivists, have been complicit in crimes against humanity. The logical positivists were certainly acquainted with crimes against humanity but in a somewhat different sense (*ibid*, p.720):

> Against the horrendous mythologies and superstitions *propagandized by the Nazis, using the old metaphysics* as a tool, these philosophers used the clarity of science to dispel non-sense and to defend common sense. Accordingly, the group was broken by the Nazis ...

The central feature of 'positivism' in geography, in the minds of its critics, appears to be the empiricism of the logical positivists. This ties in with the notion that geography in the quantitative revolution was monolithically inductivist (see above). Thus, the terms 'positivist' and 'quantitative' have come to be used more or less interchangeably by the critics, with both failing to capture the heterogeneity of the 'scientific' movement *in the discipline*. However, it is not just the empiricism of the logical positivists that the critics wish to carry over into their notion of positivism. It is the failures of logical positivism as a philosophical doctrine.

In a conversation with Brian Magee, A.J. Ayer, the man who did most to propagate logical positivism in the English-speaking world, was clear about its inadequacies (Magee 1978 p.131).

> MAGEE But [logical positivism] must have had real defects. *What do you now, in retrospect, think the main* ones were?
>
> AYER Well, I suppose the most important of

the defects was that nearly all of it was false.

The critics of positivism in geography would like to be able to claim that this observation may be extended to the foundations of the quantitative revolution and its modern manifestations. Pickes, for example, seems to think that the intellectual battle has been won by critical theorists but that the quantifiers have failed to accept defeat. He says (Pickles 1995 p.25) that for some scholars, apparently including himself,

> GIS represents a reassertion of instrumental reason in a discipline that has fought hard to rid itself of notions of space as the dead and the inert, and, as Soja (1989) has argued, to reassert a critical understanding of the sociospatial dialectic.

But this will not do as a mapping of the wider philosophical debates into a geographical context. Logical positivism has not been abandoned in favour of the critical doctrines of the so-called continental philosophers. On the contrary, it is the analytic tradition, in which logical positivism played a central part, that has come to dominate the philosophical landscape. According to Searle's essay on contemporary philosophy in the United States (*op. cit.* p.1),

> Without exception, the best philosophy departments in the United States are dominated by analytic philosophy, and among the leading philosophers in the United States, all but a tiny handful would be classified as analytic philosophers.

Magee and Ayer make a similar point at a personal level. Logical positivism may have had its day but the general view of the world implicit in it is alive and well:

> MAGEE So, a former Logical Positivist such as yourself, although you now say that most of the doctrines were false, still adopts the same general approach; and you are still addressing yourself to very much the same questions, though in a more liberal, open way?
>
> AYER I would say so, yes.

Thus, to understand the shortcomings of scientific work in geography, it is more instructive to look at the changing view of science within the analytic tradition than to turn to the philosophically eccentric positions of various critical theorists.

Returning to the Pickles quotation, one might argue that the attempts to 'reassert a critical understanding of the sociospatial dialectic' are intended to undo the very thing that logical positivism did succeed in doing – undermining the old metaphysics - but I do not want to pursue that line of argument. Rather, I want to conclude this section by asserting that the blanket attachment of the title 'positivist' to scientific work in geography does not serve to undermine the philosophical foundations of that work. Scientific geography continues to derive philosophical support from the analytic tradition, notwithstanding the demise of logical positivism, and that tradition is the dominant one in philosophy.

To summarise, the 'quantitative revolution' was neither quantitative (if that term is used to mean inductivist) nor revolutionary (in the Kuhnian sense). The heterogeneous body of work that comes under the rubric of the quantitative revolution and/or geocomputation is best described as being 'scientific'. It is not the case that the supposedly 'positivist' geography of the quantitative revolution has been weeded out by critical theorists, only to start spreading again through the development and use of GIS.

The scientific approach to geographical problems was and is firmly rooted in the analytic tradition of philosophy. Rather than turn to critical theory to understand the shortcomings of scientific geography, it is helpful to consider the changing notion of science within the analytic tradition and the changing role of computation in science.

Science and computation

It was noted above that the geography of the quantitative revolution exhibited a range of activities that gave the discipline some resemblance to the (then) standard model of a science. Specifically, geography became increasingly concerned with, on the one hand, data exploration and inductive reasoning and, on the other, model building and theory development. The prevalent notion of science owed much to the model of physics. Science was thought to be truth-seeking, cumulative, and objective. It was believed that as

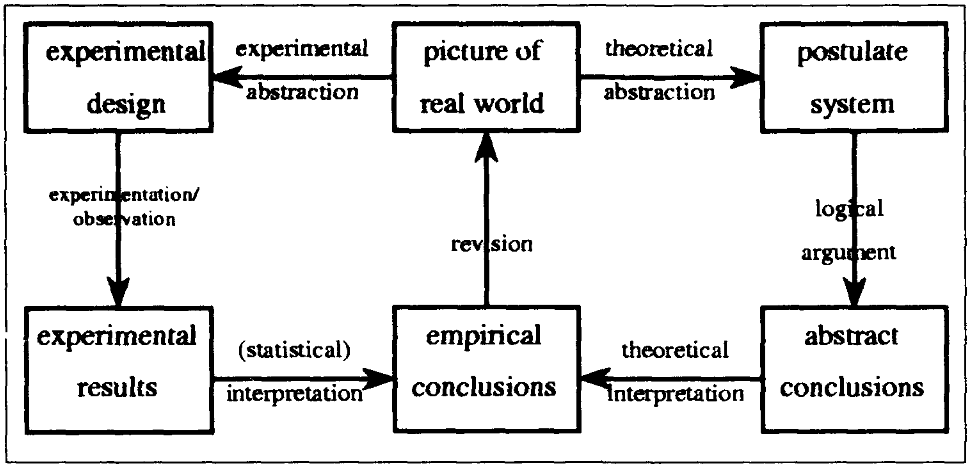


Figure 1 The scientific method in diagrammatic form

our understanding of various systems increased, so did their predictability. The process of scientific advancement was thought to consist of interrelated cycles of rationalist and empiricist endeavour (see Figure 1). Computation entered the process both in the analysis of observational and experimental data (the left hand cycle of Figure 1) and through the numerical solution of mathematical problems for which analytical techniques were inadequate (a possible strategy on the right hand side). But computation was seen primarily as a means to an end, not as part of the intellectual milieu shaping the way in which scientific problems are conceived. The social context of scientific endeavour was one of optimism about the benefits that science could bring. Consequentially, perhaps, the sociology of science was not of great interest, certainly not in a geographical context.

I want to consider some of the changes that have occurred in this view of science and its social circumstances but first I want say something about computation. Three examples should serve to illustrate the range of social and intellectual purposes to which computational devices have been put. All three examples are of significance in the history of geography.

The first is the chronometer, specifically John Harrison's four chronometers H-1 to H-4. Sobel's entertaining book *Longitude* tells of the trials (literally) and tribulations associated with Harrison's attempt to solve the problem of calculating a ship's longitude at sea. The problem was one of such importance in the early 18th century that the British Parliament, in passing the Longitude Act of 1714, offered a prize £20,000 for its solution. Two strategies came to the fore: the astronomical ideas of the scientific establishment; and Harrison's idea that it was possible to make a clock of great accuracy with which the true time could be carried from the home port. Solar observation could then be used to establish local time and the time difference used to calculate the change in longitude. The astronomers believed that no one could build a clock of sufficient accuracy. They thought that the problem would be solved by producing tables of data relating to the position of the moon relative to other celestial objects at given times and at given longitudes for years into the future. The battle, which raged through the second and third quarters of the century, provides a useful case study of the sociology of science. Sobel (1995 p.9) observes of Harrison that:

> His every success … was parried by members of the scientific elite, who distrusted [his] magic box. The commissioners charged with awarding the longitude prize – Nevil Maskelyne [the fifth astronomer royal and Harrison's principal rival] among them – changed the contest rules whenever they saw fit, so as to favour

the chances of the astronomers over the likes of Harrison and his fellow "mechanics". But the utility and accuracy of Harrison's approach triumphed in the end. His followers shepherded Harrison's intricate, exquisite invention through the design modifications that enabled it to be massed produced and enjoy wide use. Harrison's chronometers were mechanical computers dedicated to the task of measuring longitude. They are thought of as scientific instruments but are not scientific in the sense that they facilitated either the inductive or deductive processes of scientific development represented in Figure 1.

This is not true of the Varignon Frame. It can be thought of as a mechanical computer, again dedicated to a particular type of task. But the task may be thought of as belonging to the right hand side of Figure 1. The Frame computes solutions to what geographers refer to as Weberian location problems (see Wesolowsky (1993) for an interesting account of the genesis of this class of problems). That is, it provides a mechanical method for obtaining numerical solutions to a mathematical problem and, by analogy, identifies the implications of a set of assumptions about industrial location under specified initial conditions. The simplicity of the assumptions and conditions has the effect of detaching the process from the inductive, left hand side, of Figure 1; the assumptions and conditions are not capable of being true of many or any real systems so there is no sense in trying to test them. The reason why they cannot be anything but simple is, of course, the computing technology. Given the absence of an analytical solution to the general Weber location problem, a mechanical analogue computer may be used. However, as well as allowing a solution to be found, this approach limits the way in which the problem may be conceptualised.

There is an interesting parallel with the notion of market equilibrium. The idea that price and quantity in a market is determined by the intersection of supply and demand schedules is a construction rooted in 19th century computing conditions. The simultaneous solution of two equations provides answers to questions about a market that would be difficult to generate otherwise, given those conditions. But in a modern computing environment, there is not need to assume away the whole, messy, multi-agent process of market interaction. I will return to these observations later. Meanwhile, I want to consider, very briefly, some of the aspects of the changing picture of science noted above.

At the time of the quantitative revolution, one of the objections frequently raised against the use of the scientific method to study social phenomena was that it entailed a mechanical view of the world. There was some truth in this charge. Physics was the model science and mechanics is a branch of physics. Our understanding of the universe was built on a clockwork conception of the heavens. Much of the mathematics that was available, including the calculus of Newton and Leibnitz, was forged in the study of physical phenomena. And, of course, some of the approaches that were adopted were directly mechanical – like the use of the Varignon Frame. It is not too difficult to object to the employment of scientific methods in a social context when physics is the inspiration, as it was in the gravitational and thermal models of migration of Ravenstein (1885) and Hotelling (1979). But the rise of the biological sciences has altered our conception of what constitutes a science, undermining this source of objection; the intellectual distance from ecology to population geography seems less than that from physics. Indeed, as the social sciences have developed alongside the biological, the intellectual traffic has not been all one way, as it was with physics. Darwin's debt to Malthus is well know (see for example, Bronowski 1973) but it is not the only example of the biological sciences borrowing from the social; the theory of games is a more recent example of some importance (see, for example, Nowak and May, 1992).

Closely connected with the scepticism about physics as a methodological beacon, was the notion that social life does not have the predictability and, therefore, the controllability of the physical world. This belief has been undermined not so much by successes in the social arena as by the growing realisation that aspects of the physical world are fundamentally unpredictable. Interest in catastrophies and bifurcations, in fractal geometry, and in chaotic behaviour,

has spanned the scientific spectrum and this interest has helped to make it clear that, if there is a methodological cleavage between the social and physical sciences, it does not centre on predictability.

One of the other sources of this supposed cleavage is the problem of objectivity. It has long been argued that objectivity is possible in the physical world – in the study of objects – but not in the social world. But history suggests that the distinction is not so clear. The example of the longitude problem is helpful here, notwithstanding the fact that it is as much technical as scientific. The identification of the longitude problem as being worthy of study was clearly social, and the assessment of the empirical claims made by Harrison and his opponents was scarcely objective. There was as much attachment to belief in a cherished theory - and as much chicanery to sustain that belief - as might be found in any strictly social context.

The debate about objectivity shades off into the debate about truth. The revisionary metaphysics of the postmodernists is sceptical about claims to both. But when it comes to truth, the objectors have a serious obstacle to overcome. Scruton (1994 pp?) puts it this way:

> Nietzsche ... argued that there are no truths, only interpretations. But you need only ask yourself whether what Nietzsche says is true, to realise how paradoxical it is. (If it is true then it is false! - an instance of the so-called liar paradox). [GAP?] Likewise...Foucault repeatedly argues as though ... [t]here is no trans-historical truth about the human condition. But again, we should ask ourselves whether that last statement is true: for if it is, it is false...A writer who says that there are no truths, or that all truth is 'merely relative' is asking you not to believe him. So don't.

Despite counter-attacks such as this, relativism has been a mainstay of critical approaches in geography. It has taken the subject in two directions - towards a change of context and towards a change of focus.

A standard philosophical distinction is that between the context of discovery and the context of validation. Questions related to the former belong to the sociology of science; they deal with the circumstances under which particular problems and ideas have become objects of study. Questions related to the latter are methodological; they deal with the so-called logic of justification – with arguments about the reliability of knowledge claims. Philosophical concerns in geography have shifted under the influence of relativist thinking from the context of validation to the context of discovery. In 1969, Harvey's *Explanation in Geography* concentrated on methodological issues; his presumption was that there is a real world out there, which is knowable, provided certain methods are employed. Relativist dissent from this position shifted the debate to the context of discovery so that, for example, interest in Weber's theory of industrial location (such as it was) moved from the theory's propositions to its social origins and uses.

The change of focus brought about by relativist thinking has been from the general to the particular or, to use the terminology of an old debate, from the nomothetic to the ideographic. The postmodern enthusiasm for the recognition of alternative voices and the celebration of difference is underpinned by a rejection of the idea that there is a single truth, independent of the observer. This rejection relies on a rather loose usage of the term 'truth'. It may be that different individuals and groups have different perceptions of some object or phenomenon and that we cannot talk about which is the 'true' perception. But that does not mean that true propositions about what these different perceptions consist of cannot be formulated. It is important to note that this is not a repudiation of the idea that alternative voices should be heard and differences celebrated. It is a repudiation of the idea that these objectives are incompatible in principle with a scientific conception of the pursuit of knowledge. The extent to which they are compatible in practice is, at least in part, a computational issue.

Models

It should not be assumed from the foregoing argument that the notion of 'truth' is unproblematic. Indeed, in recent decades, there have been important advances in deal-

ing with this notion in both science and philosophy. In science, the dominance of Boolean logic, in which the only truth values are 0 and 1, has been reduced by the development of fuzzy logic, with its continuum of truth values (for a basic introduction with geographical references see Macmillan 1995). In philosophy, the notion of truth has been at the centre of increasingly sophisticated criticisms of realist beliefs. These criticisms have led Aronson *et al.* (1994) to mount a rescue of realism based on a re-orientation of the debate away from the truth of propositions towards the verisimilitude of models. The increasing importance of models philosophically has not been reflected in geographical work.

One of the difficulties surrounding model use in geography is that the nature of models and the purposes of model building are widely misunderstood, even amongst those who promote their use. As I have droned on at some length on these matters in other papers (see, for example, Macmillan 1989, 1996), I will confine myself here to one point. It is often said in introductions to modelling, that models involve simplifications of reality. This is true but unhelpful. First, all attempts to characterise the world, including ordinary language descriptions, involve simplifications. There is nothing peculiar, in this respect, about model building. Second, the simplicity of a model, or an ordinary language description, depends on the purposes of the author. To make this point whilst teaching I tend to pick an everyday object, like a waste bin, and ask students to describe it. As often as not, they launch into a rather complex account: 'It's a truncated cone, inverted with an open base, made of metal, painted grey, etc.'. They sometimes look puzzled when I give them my description: 'It's a waste bin'. But they see the point when I indicate the purpose of the description: 'Throw this in the inverted, truncated cone for me will you?'. My simple description is adequate for the purpose of using the waste bin. Map making is equally purposeful. The purpose of the London Underground Map is to help travellers navigate. The representation of the system is simple in order to facilitate this task – nomenclature and topology are represented accurately but nothing else is. But there are other maps of the Underground, such as those used for engineering works in the tunnels, and these attempt to represent accurately those features that are required by the engineers. The complexity of models, like that of maps, is a reflection of the purposes of their authors and users.

There are, however, technical and intellectual constraints on the achievement of these purposes. It was noted above that the Varignon Frame computes solutions to Weberian location problems. The Frame is a representation – a model – of an economic landscape. It is a 'simplification' of the landscape not because simplicity best serves the purpose of emulating the industrial location decision problem but because the computing technology will not allow greater sophistication. Similarly, but more subtly, the notion of market equilibrium embodies an intellectual constraint imposed by 19th century computational capabilities.

GIS and explanation

This brings me to the nature and use of GISystems. What are we capable of doing with this late 20th century computing technology? If one believes Taylor and Johnston, we cannot use it successfully in an explanatory context. They argue (*op. cit.* p.61) that:

> quantitative procedures, and hence GIS, ... cannot produce substantial answers to the question 'Why?'

They base this view on Sayer's (1984) notion that mathematics is an acausal language. I have taken issue with this claim before (Macmillan, 1989). If we regard 'cause' as 'sufficient condition' (see, for example, Hospers 1967 p. 279-320), then a set of mathematical relations with an appropriate empirical interpretation can be construed causally. This is precisely how the causal explanations of the physical sciences are formulated: a set of equations, say, represents a set of law-like generalisations; a set of parameter value assignments constitutes a set of condition statements; and a solution statement represents the statement of the event or condition to be explained.

GISystems *should* be used in producing substantial answers to the question 'Why?'. They allow a representation of spa-

tial systems which is substantially better than that embodied in the Varignon Frame. They provide a less-constraining computational environment. They certainly do not provide a *non*-constraining environment and some of the critical comments that have been made about data-led GIS work may be thought of as highlighting some of the constraints that undoubtedly operate.

Taylor and Johnston further question the possibility of using GIS for explanatory work by arguing (*ibid.* p.57) that

> The original 'quantifiers' attempted to ... [develop] deductive theory but ... it is just this aspect of quantitative geography that has been severely castigated by GIS proselytizers ...

That is a fair comment, taken in isolation, although it is a little surprising to find the proselytizers called on in support of a case that is largely directed against them. But as a line of argument it is not persuasive. The fact that Openshaw sees science as consisting only of the inductive half of Figure 1 does not make it so. And the suspicion that Openshaw can see more out of his one methodological eye than many can with two does not alter this conclusion.

It is certainly the case that much GIS work has been data-led and that a good deal of it has been applied. But it is also true that there has been a fair amount of theoretical endeavour. Goodchild (1995 p.46) notes that

> An environmental modeler will likely write his or her model in source code, typically FORTRAN or C, but may well maintain a GIS, linked to the modeling system, to preprocess data, and to analyze and present the model's results. This type of GIS use probably characterizes the majo[illegible] of efforts in environmental simulation modeling...

Theoretical endeavour of this kind bridges the gap between pure and applied geography, to which Taylor and Johnston allude. That gap, as indicated above, is largely computational in origin. A rich system of conditions, on which law-like statements can operate, allows theoretical ideas to be applied.

Social change

There is a greater continuity here than Taylor and Johnston would allow. What they see as a 'tension' between pure and applied work in the quantitative revolution does not look like that to me. For my own part, starting work in geography too late to be a revolutionary, theory seemed to be a necessary pre-requisite for application. Indeed, the thing that was applied *was* the theory. I became interested in theory development because of my interest in applications and many others did the same. Of course, the social climate was one in which it was thought desirable to provide scientific support for rational decision making in the public interest. Much computational model building was predicated on this idea. But societies change.

The culture of the times, in many parts of the world, swung against what might be called the planning perspective. From the right, it was not just planning and the social democratic notion of market intervention that came under attack - it was the notion of society itself. From the left, the supposed irredeemability of capitalism led to a similar conclusion – the idea of rational decision making in the public interest was a snare and a delusion. But, as I have just said, societies change.

In Britain, in much of Western Europe, in the U.S. (arguably), and in many other places, the intellectual leadership of the right has waned. At the same time, the dramatic collapse of communism has done little to further the claims of the left. Geography as a discipline has become somewhat eccentric in its continuing interest in Marxist thinking - much of the rest of the academy has moved on. To be sure, the new world is not the same as the old, either materially or intellectually. But the old idea that science can serve society, and serve in the study of society, has re-emerged, battered but unbowed.

Geocomputation

Where does this leave us with regard to the nature of geocomputation? I don't propose to dwell on what it consists of historically or currently but I will venture an opinion on what it could or should be. The foregoing argu-

ments suggest that it should be a set of activities, conducted in or around a computationally sophisticated environment, in which the geographical sciences are developed and applied. Taking GIS to be the paradigmatic example of a computationally sophisticated environment, this means that we should be using GIS for theory development both inductively and deductively. That is, we should use GIS in an inferential mode but we should also be concerned with building models in a GIS environment – an activity that is theory-led rather than data-led. Indeed, GISystems should become the laboratories within which the two scientific cycles of Figure 1 interact fully for the first time in a geographical context.

This is not to say that application should be neglected. Theory and application should be related cyclically in what might be thought of as an orthogonal relationship to that shown in Figure 1. Theory should inform application and application should inform theory. In both cases, verisimilitude should be a watchword, although there should be an economy of design appropriate to the purposes of the exercise.

Clearly, geocomputational exercises should have explicit purposes and they should be conducted in the knowledge that those with whom we interact have their own purposes, including those that supply data and consume advice. Also, the form in which advice should be offered is of considerable concern. In applied work, we should not behave as if we were producing a product for a consumer, where the product consists of a single forecast and an optimal prescription based on that forecast. It would be more consistent with our contemporary understanding of science to build a model with which users can play, on the understanding that it can yield useful insights about real decision problems but that those insights are limited by the verisimilitude of the model (see Macmillan 1996). In theoretical work, we should take up our own purposes, the traditional purposes of the academy. For those of us concerned with society, we should be prepared to assert that our purpose is to understand, however hard that may be.

As for the critics, they might well claim that this is a pious hope, given the history of what they might see as data-led, theory-free, ethically neutral work in GIS and related fields. I prefer to think of it as a challenge to a new generation to see that the promise of geocomputation is fulfilled.

References

Aronson, J.L., Harré, R., and Way, E.C. (1994) *Realism Rescues: how scientific progress is possible* Duckworth, London

Bronowski, J. (1973) *The Ascent of Man* Book Club Associates, London

Haggett, P. and Chorley, R.J. (Eds.) (1967) *Models in Geography* Methuen

Harvey, D. (1989) From Models to Marx: Notes on the Project to 'Remodel' Contemporary Geography. In B.Macmillan (Ed.) *Remodelling Geography* Basil Blackwell Ltd., Oxford.

Harvey, D. (1969) *Explanation in Geography* Arnold

Hospers, J. (1967) *An Introduction to Philosophical Analysis* Revised Edition. Routledge and Kegan Paul, London.

Hotelling, H. (1979) A mathematical theory of migration. Environment and Planning A, 10(11), pp. 1233-9.

Kuhn, T. (1962) *The Structure of Scientific Revolutions*. University of Chicago Press, Chicago.

Macmillan, W. (1989) Quantitative Theory Construction in Human Geography (1989). In In B.Macmillan (Ed.) *Remodelling Geography* Basil Blackwell Ltd., Oxford.

Macmillan, W. (1995) Fuzziness revisited. Progress in Human Geography, 19.3.

Macmillan, W. (1996) Fun and games: serious toys for city modelling in a GIS environment. In P.Longley and M.Batty (Eds.), Spatial Analysis: Modelling in a GIS environment, Longman Geoinformation.

Magee, B. (1978) *Men of Ideas: some creators of contemporary philosophy* BBC, London.

Nowak, M.A. and May, R.M. (1992) *Evolutionary Games and Spatial Chaos* Nature, Vol. 359, pp 826-829

Pettit, P. (1993) The contribution of analytic philosophy. In R.E.Goodin and P.Pettit (Eds.) *A Companion to Contemporary Political Philosophy* Basil Blackwell, Oxford.

Pickles, J. (1995) Representations in an electronic age. In J.Pickles (Ed.) *Ground Truth: the Social Implications of Geographic Information Systems* The Guilford Press, New York.

Ravenstein, E.G. (1885) The Laws of Migration. Journal of the Royal Statistical Society, 48, pp. 167-235.

Sayer, A. (1984) *Method in Social Science: a realist approach* Hutchinson, London.

Scruton, R. (1994) *Modern Philosophy: an introduction and survey* Mandarin, London

Searle, J.R. (1996) Contemporary Philosophy in the United States. In N.Bunnin and E.P.Tsui-James (Eds.) *The Blackwell Companion to Philosophy*. Blackwell Publishers Ltd., Oxford.

Sobel, D. (1995) *Longitude* Fourth Estate, London

Soja, E. (1989) *Postmodern Geographies: the reassertion of space in critical social theory* Verso, London.

Solomon, R.C. (1997) *Introducing Philosophy: a text with integrated readings* 6th Edn. Harcourt Brace, Orlando.

Taylor, P.J. and Johnston, R.J. (1995) GIS and Geography. In J.Pickles (Ed.) *Ground Truth: the Social Implications of Geographic Information Systems* The Guilford Press, New York.

Wesolowsky G.O. (1993) The Weber Problem: history and perspectives. Location Science 1.1, pp. 5-23.

A hybrid rule-object spatial modelling tool for catchment analysis

P G Luckman, R D Gibson[1], and R R Chamarti
Landcare Research, Private Bag 11 052,
Palmerston North,
e-mail luckmanp@landcare.cri.nz

*Presented at the second annual conference of GeoComputation '97 & SIRC '97,
University of Otago, New Zealand, 26-29 August 1997*

Abstract

Sustainable land management requires understanding of the cumulative effects of current and likely future land use patterns.A modelling shell (LAMS, Land Management Simulator) has been developed to allow exploration of how on- and off-site effects develop in time and space. LAMS is tightly integrated with the Arc/Info GIS, and models can freely access spatial data, execute GIS spatial operations and manipulate the spatial display.An application of LAMS to land use change in New Zealand erodible hill country is discussed.

Introduction

Sustainable management of productive hill country catchments is a key concern for New Zealand resource management agencies, communities, central government, and industry (Ministry for the Environment, 1996).While there continues to be concern about sustainability of pastoral land use in many hill country areas, there is also a need to ensure that emerging land use patterns provide an appropriate balance between possible detrimental effects, such as reduced water resources, and beneficial effects such as reduced soil erosion.

Resource management agencies and communities need information on land use effects in "large" catchments. Management questions relate not only to the magnitude and timing of effects, but also to priorities for data collection. However,"large" catchments present difficulties in that they are not only physically large compared to small research catchments, but also highly heterogeneous in terms of both the land resources and the processes operating within them. Consequently, predicting the behaviour of such systems represents a challenge for modellers and analysts (Kalma and Sivapalan, 1995).

Although there is a need for greater understanding of the processes operating within catchments, providing practical support for catchment analysis requires appropriate tools for integrating and applying knowledge about spatially distributed systems. Consequently, there is interest in combining knowledge engineering tools with geographic information systems to provide comprehensive spatial modelling technologies for addressing catchment analysis problems (e.g. Mackay *et al.*, 1993; Lam and Swayne, 1993; Reynolds *et al.*, 1996).

Our goal has therefore been to develop tools which support both building and applying process-based and interpretive models for predicting the behaviour of hill country catchment ecosystems. Starting with sedimentation analysis, we have developed a modelling tool (LAMS, Land Management Simulator) for investigating catchment land use effects. This paper describes the design of LAMS and its application to a lake sedimentation problem associated with pastoral land use in an erodible North Island catchment in New Zealand.

Computer support for catchment analysis

Analysis of land use effects can proceed in three stages

[1]Now at Paradigm Technologies, Wellington, New Zealand.

(Bartell, 1996). In the first stage, the landscape is scanned for the occurrence of risk-generating situations - which will generally be characterised in terms of associations of land use (or management) and land type. Secondly, potential outcomes of these situations need to be formulated, in consultation with the community. Finally the likelihoods of these outcomes must be determined, preferably in terms of risk probabilities.

Society's tolerance of undesirable land management patterns depends on their cumulative effects (Cocklin *et al.*, 1992). Effects may accumulate insidiously over time; they may be the total effect of risk-generating situations upstream or the end result of a cascading sequence of indirect effects; or they may simply consist of a collection of diverse effects. Accordingly a tool for catchment analysis needs to provide facilities to build and link many types of models of varying sophistication, each model reflecting the state of knowledge as well as the availability of data to apply it.

Saarenmaa *et al.* (1994) have shown that decision support and analysis for natural resource management is most effectively provided if the system being modelled can be represented as a set of objects (Coad and Yourdon, 1991) which correspond to real world objects. This "computational framework" provides the foundation for a variety of models, leading ultimately to a library of compatible domain-dependent tools for the particular resource management problem area.

For catchment analysis, the problem of scale has inhibited agreement on the content and structure of such an object model (Kalma and Sivapalan, 1995). Modellers typically have difficulty scaling up from sound understandings of surface and subsurface flow to models which accurately predict the hydrologic behaviour of whole catchments. For example, preferential flow pathways such as tension cracks, fissures or shear zones in unstable hillslopes are usually not considered in hydrologic models based on the differential equations for flow of water in porous media.

Notwithstanding this problem, there are fundamental concepts (or objects) which underpin catchment analysis which addresses sustainability questions (Naiman *et al.*, 1992). For example, catchments comprise subcatchments linked by stream segments. Land within the catchment consists of geomorphological units reflecting surface morphology, regolith type, geology and erosion processes. Soil classes and properties can be inferred from geomorphology using soil-landscape models. The "representative elementary area" and "hydrological response unit" are similar discrete area concepts used to model catchment hydrology (Wood *et al.*, 1988, Flugel, 1995). Further, sediment and nutrient loadings of surface and subsurface water, and the chemical transformations of the solutes, are determined by the ways in which flow pathways intersect with these soil or geomorphological units.

These and other concepts (including those which underpin modelling of socio-economic factors) potentially provide the basis for an object-oriented, spatio-temporal catchment modelling tool which can be used at a variety of scales. Because of the clarity and ease of explanation of simple rule-based interpretive models, and because resource management scientists who are not programmers need easily accessible modelling aids, rule-based knowledge representation is also required (Carrico *et al.*, 1989). Essential requirements of such a modelling tool are that it should be easy to modify and extend models to reflect the issues of concern to different communities, and that the tool provides efficient access to spatial data in a form which is easily maintained and verified.

LAMS modelling framework

Overview

The essential components of the Land Management Simulator catchment analysis system, similar in concept to that described by Fedra (1995), are shown in Figure 1. The system contains database, modelling and geographic information system components which can be accessed via graphic development and application interfaces. A core feature is a simulation manager which evolves land use and land cover patterns through time, and applies models to predict effects and changes in risk levels. *LAMS* uses both object-

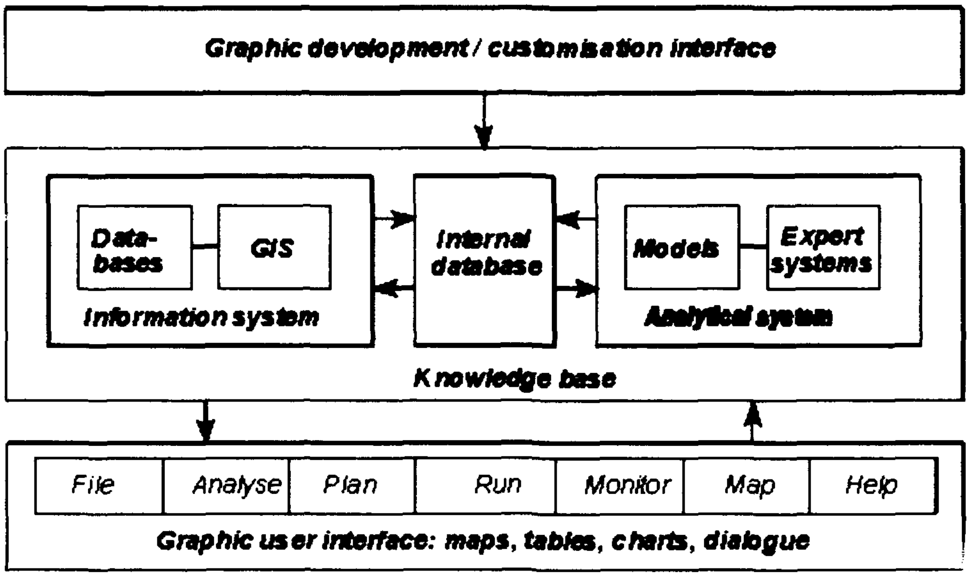


Figure 1 Architecture of the catchment modelling and analysis system

oriented and rule-based knowledge representation techniques.

LAMS has been developed on a Sun workstation. We have used the Smart Elements knowledge-based development environment from Neuron Data in combination with ESRI's Arc/Info geographic information system. Smart Elements integrates a hybrid rule / object-oriented expert system shell (Nexpert *Object)* and Open Interface, a cross-platform Graphical User Interface (GUI) developer (which may assist development of a PC version of *LAMS*). Overall control and model management are handled within Nexpert *Object*. The ease with which the flow of control and the hierarchy of data structures can be viewed within Nexpert facilitates understanding. Some simulation modelling is coded directly in C, for greater efficiency. The GUI development features of Smart Elements have been used to construct an interface (Figure 2) involving data entry and output screens in the form of editors, a network browser, charting and mapping capabilities and textual reports. Most GUI development has been coded in C, although Smart Elements provides scripting support for some GUI elements. *LAMS* can read from and write to ARC/INFO databases through a set of C routines, and can control the GIS through commands issued through an Inter-Application Communication (IAC) connection (ESRI, 1995).

Representing the spatial domain

We model the stream channel network within the catchment as a set of stream segments or reaches. Each of these is an object, inheriting attributes and operations from the stream segment class. A small number of local subcatchments (LSCs), the smallest catchment unit represented, drain into each stream segment. These local subcatchments may provide point or linear water sources to the stream segments, depending on whether they are defined around particular streams or whether they simply drain into the stream segment over part or all of its length. The total catchment for a stream segment is then the collection of LSCs for the segment and all upstream segments.

The fundamental modelling unit is the land response unit or LRU, similar in concept to both the landscape response unit in landscape ecology (Perez-Trejo, 1993), and the hydrological response unit (Flugel, 1995). We define contiguous areas of land with a common manager as socio-eco-

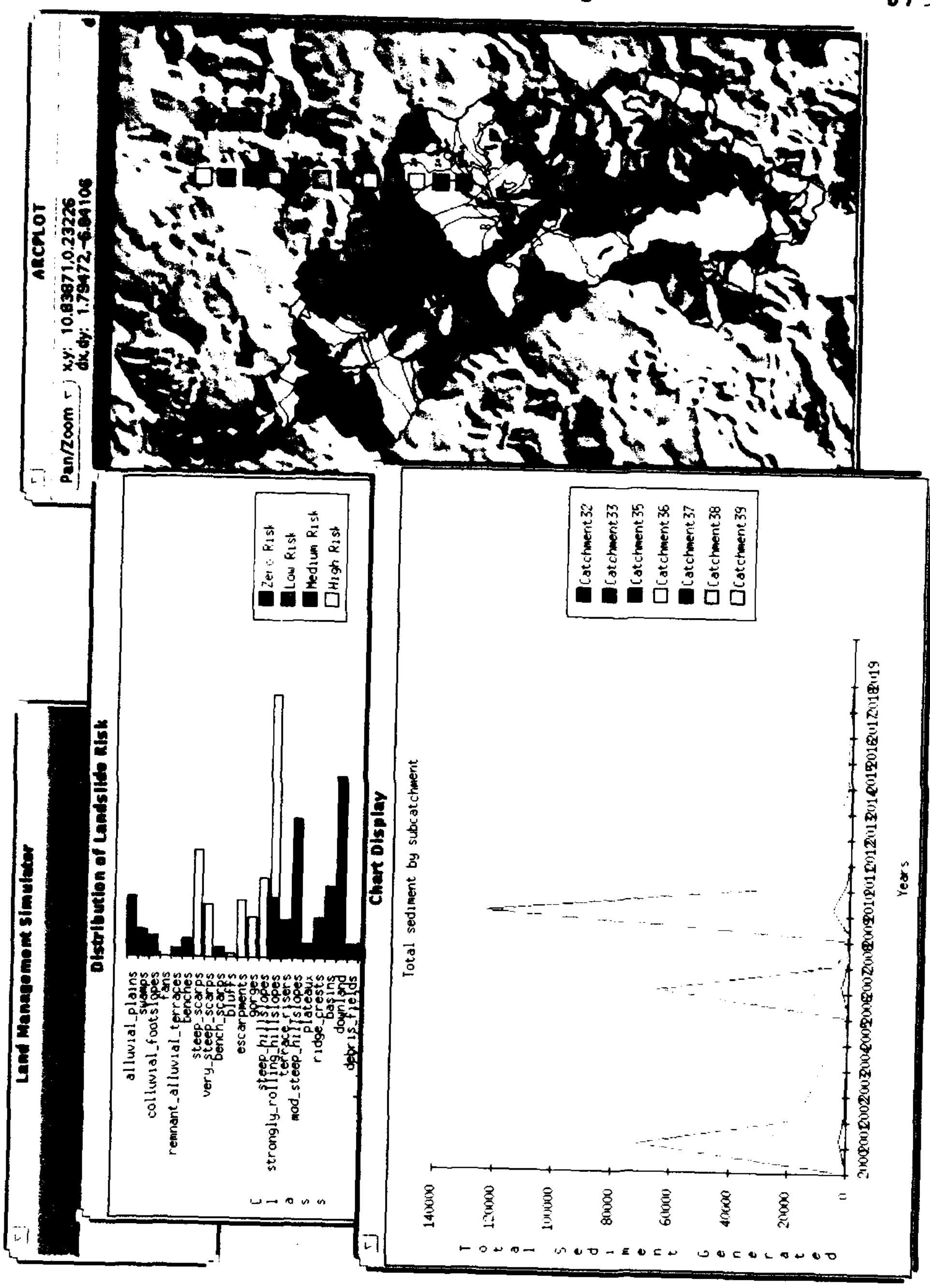


Figure 2 The Land Management Simulation graphic user interface

nomic units (SEUs), and describe the terrain through a set of geomorphic land classes (GLCs), which are available as mapped polygons or as a non-aggregated classified raster image. We then define an LRU as a conceptual unit comprising all land within a given LSC and SEU which belongs to the same GLC. Land management may vary within the LRU - which is modelled as a set of land use (or management) units (LUUs) (*Figure* 3). The *resulting* class-object hierarchy for the catchment ecosystem is represented in Figure 4, following the notation of Coad and Yourdon (1991). The object model facilitates rule-based reasoning about the system or selected parts of it, while direct representation of connectivity and parent-component relationships, in addition to classifications, supports routing of messages to appropriate objects.

Land use and vegetation change

Changes in land use are modelled using land use *transition* rules. These rules specify when and where land use change will occur, and the nature of the land use transition. The rules are currently deterministic, but could be stochastic, reflecting specified levels of uncertainty about land manager decisions or changes in land ownership (Dale *et al.*, 1993, Lee *et al.*, 1992). Data describing spatial and non-spatial pre-conditions for change, and the changes which occur, are captured on an editor screen. Each transition rule is attached as a method to an object in a class of "land use change rules".

The condition lists of rules take into account the terrain class (GLC), the position in the catchment (subcatchment or local subcatchment), the ownership, and the existing land use. Factors which motivate the land use change are treated implicitly with this representation. More complex rules which consider factors such as the state of neighbouring properties, economic indicators, or whether there has recently been a major erosion event, can be created using the graphic expert system development interface directly. Collectively, groups of land use transition rules specify land use scenarios.

Changes in land cover also occur as a result of maturing vegetation or natural succession. We employ the concept of a vegetation phase; land not in productive use follows different succession patterns (sequences of phases) under different conditions. We associate a phase sequence or succession model with each area which is removed from productive use. Each phase and its associated attributes is represented as an object within a class of vegetation phases. Each land response unit has an attribute describing the anticipated sequence and timing of phases, in case parts of the LRU are withdrawn from productive use, or already contain areas of scrub or regenerating indigenous vegetation. Succession models are allocated interactively to spatially-defined classes of LRU's.

Application to sedimentation analysis

Many environments in New Zealand are susceptible to

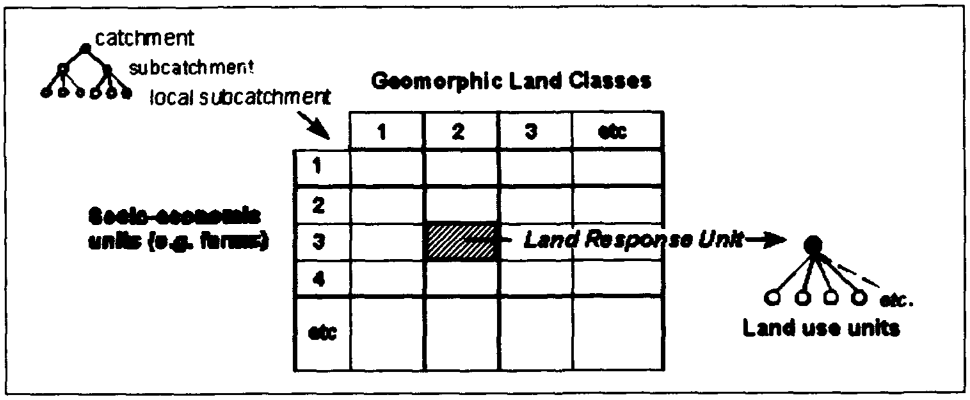


Figure 3 Basic modelling units employed in the catchment modelling analysis system.

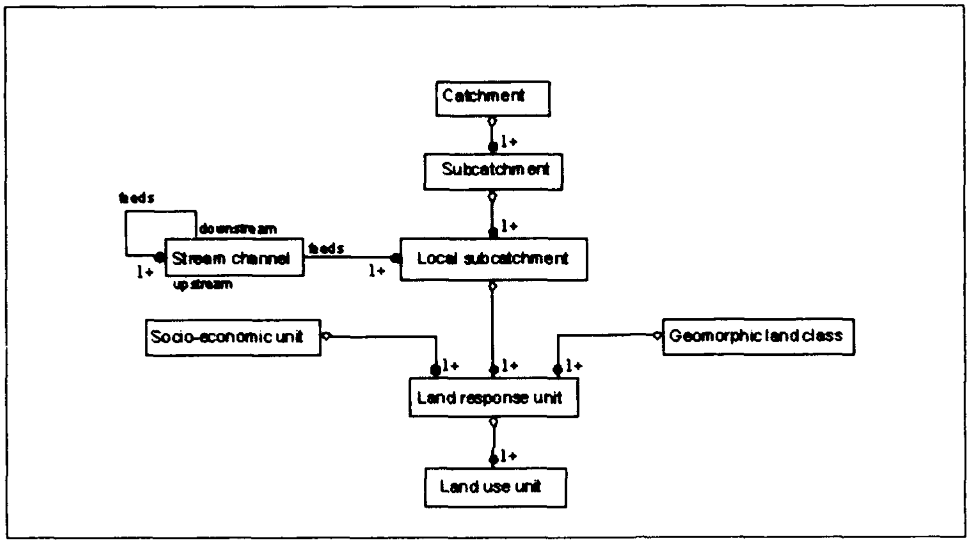


Figure 4 Class hierarchy for representing spatial and conceptual relationships within catchment

soil erosion. The resulting sedimentation can lead to flooding, loss of habitat, and reduced water quality. One example is the 32 km² Lake Tutira catchment in northern Hawkes Bay (Trustrum and Page, 1992). There, resource managers are interested in how land use change within the catchment will affect lake sedimentation which has proceeded at high rates since clearance of indigenous vegetation (Page *et al.*, 1994a). Resource managers have also expressed interest in the possibility of designing land use changes so that the catchment can withstand rainstorms of a specified magnitude or frequency without causing significant lake sedimentation.

The approach we have adopted is to simulate, on a yearly basis, the projected changes in land use or land management within the catchment, and use (at this prototype stage) very simple empirical models to suggest possible effects. In future, we anticipate elaborating these models and incorporating interpretive models to test for significance of effects and to explore indirect effects.

The catchment is subjected to a sequence of annual maximum rainstorm events, which are either specified by the user or selected randomly from an extreme value distribution. A linear empirical response model is used to compute the amount of landsliding on susceptible slopes (where slopes and rainfall exceed empirically-determined thresholds), assuming pastoral land use under "standard" management. The amount of erosion is then adjusted empirically to take account of factors such as land management, land cover, age of trees, and the available soil resource.

Chronic erosion and sedimentation delivery to streams is modelled as empirically assessed annual transfers between landform components (Reid and Trustrum, *in prep*). For each land use unit (LUU), erosion processes and sediment transfer rates for pastoral land use are inherited directly from GLCs (Figure 5). LUUs inherit methods from land use classes which allow these rates to be adjusted for the nature of the land cover.

Risk quantification requires determining the probability, following a storm of given magnitude, that more than x mm of sediment accumulates in Lake Tutira. Our approach is firstly to assess the probability under the current land use regime. While for other sedimentation problems this "current risk" might be assessed differently, for Lake Tutira we were able to use an empirical log-linear relationship between storm rainfall and the thickness of sediment deposited in the lake, obtained by an analysis of lake cores

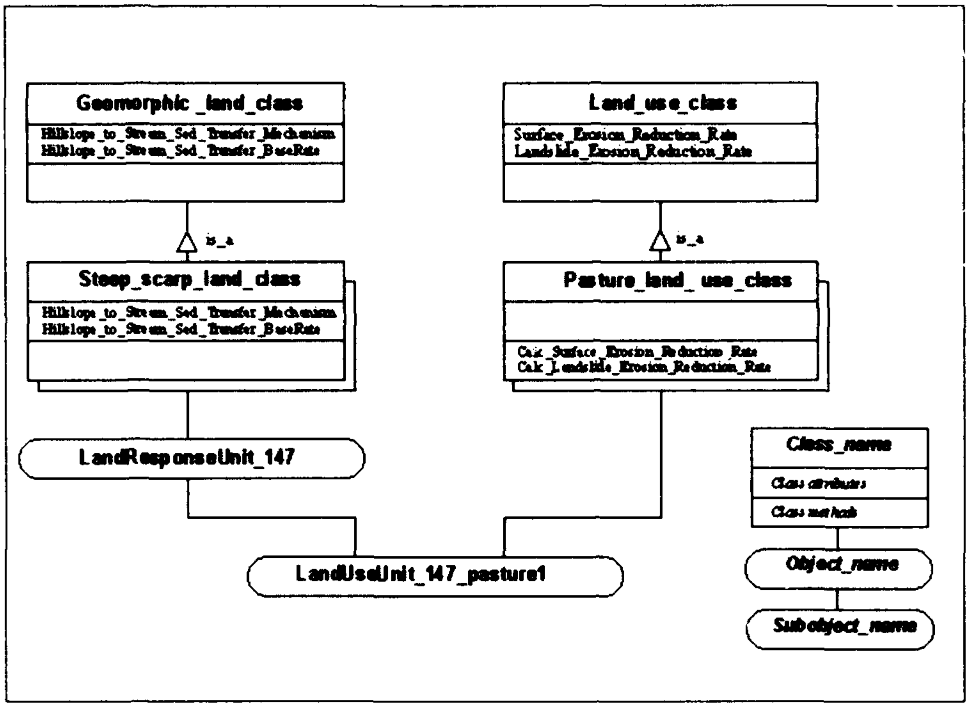


Figure 5 Inheritance of methods and attributes for computation of land use-adjusted sediment transfers

(Page *et al.*, 1994a). We then revise this probability, taking into account the likely effects of changes in land management on erosion and sedimentation processes.

To determine the effect of land use on sedimentation we first establish the major processes by which sediment is generated (by erosion or remobilisation of sediment in temporary storage) and reaches the lake, for the storm rainfall in question. For this we employ expert-derived curves giving bounds on the contribution of different erosion processes to the total volume of sediment reaching the lake, as a function of storm rainfall. A sediment budget for Lake Tutira catchment has been evaluated by Page *et al.* (1994b). Landsliding contributes most of the sediment for the large rainfall events, but the bounds are further apart for the smaller events for which processes such as streambank erosion and channel erosion can become significant.

During simulation of land use change we "monitor" the change in state of key sediment sources. Having identified the principal mechanisms responsible for the sediment delivered to the lake for a rainstorm of the size in question, for current conditions, we search the areas of land use change upstream to establish the list of land use units subject to these processes. We then determine the extent to which vulnerability to the contributing mechanism has been affected by land use change. The (increased or decreased) percentage change is computed by applying empirical factors deduced from available land use impact information. Depending on the mechanism under consideration, this requires separate evaluation of a range of factors which can affect sediment delivery. While we have not made use of hydrological models to date, we envisage these will be useful when we attempt a more rigorous analysis of the effects of land use on sediment delivery. For example, resolving whether factors such as changes to peak flows and rainfall interception rates could significantly affect the importance of individual sediment supply

mechanisms can become the focus for separate knowledge-bases.

To establish the impact of land use change on landslide erosion, we firstly search for landslide-prone areas which have undergone land use change. Where land use has changed to forestry, the land's new vulnerability to landsliding at each location is computed from the age of the trees, and from the time which has lapsed since the end of the previous harvest cycle, after the first rotation. For areas which are undergoing succession-driven changes in vegetation cover, changed vulnerability to landsliding depends on landslide-inhibiting characteristics of the land cover which are stored with the vegetation phase objects.

Worst case and best case scenarios for sediment generation are constructed using a linear programming algorithm (Winston, 1991) which determines bounds on the change in sediment delivery. Worst cases occur when the sediment sources which would be most affected by the land use change contribute the least to the sediment reaching the lake. An appropriate weighting of worst and best case reductions (e.g. the average reduction) in sediment delivery caused is used to compute a revised risk probability which is classified using simple rules. The user can then run the model for a number of years and determine when sedimentation risk becomes acceptable.

The overall logic of this analysis is represented using rules to help make it more visible (through the rule network viewer) and easily understood. Evaluating conditions and performing actions associated with these rules involves computation, queries to the internal (object) database, and further rule-based inference. The rule-set achieves spatial reasoning though queries which exploit knowledge of the upstream-downstream relationships made explicit in the catchment representation.

Conclusion

The goal of this research has been to develop and apply an object-oriented framework to support analysis of land use effects in hill country catchments. We have developed an object-oriented data model which now forms the basis for the *LAMS* catchment modelling tool. This tool, which has been built by tightly linking knowledge engineering and geographic information systems components, has been successfully applied to analysis of lake sedimentation risk. Both the underlying object model and the modelling approach used for sedimentation analysis have potential for application to problems relating to catchment hydrology, stream water quality, or to valued environmental components such as spawning grounds for fish. In future we anticipate using and further testing the *LAMS* conceptual model by developing knowledge-bases and models to address a variety of catchment management issues.

Acknowledgements

Funds for this research have been provided by the New Zealand Foundation for Research, Science and Technology under Contracts CO9306 and CO9612. The authors acknowledge the additional programming support of Chris Scott, Landcare Research, Palmerston North. Ideas about evaluating sedimentation effects of land use have been developed through discussions with Dr Leslie Reid, of USDA Forest Service at Arcata, California. Data for the Tutira case study was supplied by Noel Trustrum and Mike Page, Landcare Research, Palmerston North.

References

Bartell, S.M. 1996. Ecological / environmental risk assessment: principles and practices. Chapter 10 *in* Kolluru, R.V.; Bartell, S.M.; Pitblado, R.M. and R.S. Stricoff. *eds.* Risk Assessment and Management Handbook for Environmental, Health and Safety Professionals. McGraw-Hill.

Carrico, M.A.; Girard, J.E.; Jones, J.P. 1989. Building Knowledge Systems: Developing and Managing Rule-Based Applications. Intertext Publications, McGraw-Hill, 335p.

Coad, P.; Yourdon, E. 1991. Object-oriented Analysis and Design. Prentice-Hall.

Cocklin, C., Parker, S.; Hay, J. 1992. Notes on cumulative environmental change. I: Concepts and Issues. *Journal of environmental management* 35: 31-49.

Dale, V.H.; O'Neill, R.V.; Pedlowski, M.; Southworth, F. 1993. Causes and effects of land use change in central Rondônia, Brazil. *Photogrammetric engineering and remote sensing* 59: 997-1005.

ESRI. 1995. ARC/INFO Version 7.0 Online Documentation. Environmental Systems Research Institute, Redlands, California.

Fedra, K. 1995. Decision support for natural resource management: models, GIS and expert systems. *AI Applications* 9: 3-19.

Flugel, W-A. 1995. Delineating hydrological response units by geographic information system analyses for regional hydrological modelling using PRMS/MMS in the drainage basin of the River Bröl, Germany. *Hydrological Processes* 9: 423-436.

Kalma, J.D; Sivapalan, M. 1995. Scale Issues in Hydrological Modelling. Wiley. 489 pp.

Lam, D.C.L; Swayne, D.A. 1993. An expert system approach to integrating hydrological databases, models, and GIS: application of the RAISON system. Proceedings of the Vienna HydroGIS '93 Conference. Application of GIS to Hydrology and Water Resources. *IAHS Publication* 211: 23-33.

Lee, R.G.; Flamm, R.; Turner, M.G.; Bledsoe, C.; Chandler, P.; DeFerrari, C.; Gottfried, R.; Naiman, R.J.; Schumaker, N. 1992. Integrating sustainable development and environmental vitality: a landscape ecology approach. Pages 499-521 *in*: Watershed management: balancing sustainability and environmental Change, Naiman, R.J. *ed*. Springer-Verlag.

Mackay, D.S.; Robinson, V.B; Band, L.E. 1993. An integrated knowledge-based system for managing spatio-temporal ecological simulations. *AI Applications*, 7: 29-36.

Ministry for the Environment. 1996. Sustainable land management: A strategy for New Zealand. 16 pp.

Naiman, R.J.; Beechie, T.J.; Benda, L.E.; Berg, D.R.; Bisson, P.A.; MacDonald, L.H.; O'Connor, M.D.; Olson, P.L.; Steel, A. 1992. Fundamental elements of ecologically healthy watersheds in the Pacific Northwest Coastal Ecoregion. Pages 127-187 *in*: Naiman, R.J. *ed*. Watershed Management: Balancing Sustainability and Environmental Change. Springer-Verlag.

Page, M.J.; Trustrum, N.A.; DeRose, R.C. 1994a. A high resolution record of storm-induced erosion from lake sediments, New Zealand. *Journal of paleolimnology* 11: 333-348.

Page, M.J.; Trustrum, N.A.; Dymond, J.R. 1994b. Sediment budget to assess the geomorphic effects of a cyclonic storm, New Zealand. *Geomorphology* 9: 169-188.

Perez-Trejo, F. 1993. Landscape response units: process-based self-organizing systems. Pages 87-98 in: Landscape Ecology and Geographic Information Systems (R. Haines-Young, D.R. Green, S. Cousins, editors), Taylor and Francis.

Reid, L.M.; Trustrum, N.A. (*in prep.*) Sediment budgets for land management planning: opportunities and challenges. To be submitted to *New Zealand Geographer*.

Reynolds, K., Cunningham, P.; Bednar, L.; Saunders, M.; Foster, M.; Olson, R.; Schmoldt, D.; Latham, D.; Miller, B.; Steffenson, J. 1996. A knowledge-based information management system for watershed analysis in the Pacific Northwest U.S. *AI Applications* 10: 9-49.

Saarenmaa, H.; Perttunen, J.; Vakeva, J.; Nikula, A. 1994. Object-oriented modelling of tasks and agents in integrated forest health management. *AI Applications* 8: 43-59.

Trustrum, N.A.; Page, M.J. 1992. The long term erosion history of Lake Tutira watershed: implications for sustainable land use management. Proc. International Conf. On Sustainable Land Management, Napier, Hawkes Bay, New Zealand, 212-215.

Winston, W.L. 1991. Introduction to Mathematical Programming Applications and Algorithms. PWS-Kent Publishing Co., 794 pp.

Wood, E.F.; Sivapalan, M.; Beven, K.J.; Band, L.E.. 1988. Effects of spatial variability and scale with implications to hydrologic modeling. *Journal of hydrology* 102: 29-47.

Simulation Tools for Transparent Decision Making in Environmental Planning

Cedric Grueau, Armanda Rodrigues
Centro Nacional de Informação Geografica
R. Braamcamp, 82 - 5º Esq
1250 Lisboa - Portugal
Tel: + + 351 (1) 3860011 - Fax 351-1-3862877
http helios.cnig.pt
e-mail:(cedric;armanda)@cnig.pt

Presented at the second annual conference of GeoComputation '97 & SIRC '97, University of Otago, New Zealand, 26-29 August 1997

ABSTRACT

The planning and territorial management process is often disparaged and subjected to criticism based on the way environmental decisions are taken. The lack of transparency and the high technical level of the Environmental Impact Assessment process do not ease public agreement. The increasing development in Geographical Information technologies has helped the construction of GIS-based Spatial Decision Support Systems (SDSS) enabling 'multi-purpose planning'. The SDSS example presented below shows the potential for integration of several levels of involvement around an open platform aiming at a more scientific and shared decision in environmental planning. However, the development of this environmental SDSS has lead to the identification of a major need for an engagement effort towards the structuring and normalisation of the information to be created and published. It has become necessary to develop methodologies that will enable the systematisation, modelling, quantification and qualification of geographic space. This paper proposes that the definition of minimal geographic elements and the conceptualisation of geographical space into such description components leads to the creation of structures which allow for the thorough application of spatial analysis in environmental planning.

1 INTRODUCTION

Official forecasts indicate that, without drastic changes in policies, environmental quality will deteriorate over the coming years. The pressure on the environment will impair its potential to provide functions to society such as supply of drinking water, forestry and recreation. It is of major importance that planning activities become primarily based on environmental concerns. Moreover, Portugal has, in the last 10 years, dealt with strong economical concerns to achieve the now forthcoming challenges dictated by the european context. The resulting environmental pressure led to a major necessity in the definition of tools that could help decision makers scientifically integrate environmental values into the planning process and, at the same time, keep this integration transparent and understandable to the public.

The project presented focuses on the possibility of simulating the effects of human actions and land use transformations on an interactive basis. It is based on land use characterization through the association of hazard effects with types of land use. This paper describes the inception of the project, its first steps and current state, with new methodologies and technology being fed into it. It also presents new developments related to representation models which, we argue, will definitely improve its performance as a decision-making tool.

2 A DECISION SUPPORT SYSTEM FOR MUNICIPAL ENVIRONMENTAL PLANNING

The basis for this work was the SIGLA project (GIS Simulation of integrated environmental indicators). It was based on a decision model primarily conceived with the objective of providing a standard approach for planning evaluation methodologies and a new basis for normative procedures. Its initial funding (provided by The Portuguese Environmental General Direction) aimed to get results as tools for the development and assessment of new rules in planning activities. Moreover, it was necessary to provide planners with entertaining experimentation tools for the calculation of alternative planning scenarios (Janssen, 1991). In this way, the conception and application of normative could be simulated and evaluated at the desktop. The model is structured according to a simulation/evaluation approach. Evaluation perspectives are provided at three levels: The expert level, the municipal level and the public level. The expert level requires the intervention of a team of planning experts to define a system of dependencies between the model components and a set of rules for normalisation in the definition of weights. At the Municipal level a team of technicians defines the set of weights that implement their municipality policy and perspective, following the normalisation rules defined at the previous level. The public can demonstrate its preoccupations (Shiffer, 1992) by suggesting modifications to the perspective applied by the municipality. This provides the model transparency component often lacking in the decision process by allowing non-technical users to interact with its implementation, modify its criteria and evaluate the result of changes. The project was instanciated in a agent-based SDSS providing decision elements from simulation of changes. This system supports evaluation, simulation of changes according to the evaluation perspective, integration of judgement with methods and data and processing of all relevant information (Janssen, 1991). Therefore, the conceptual definition includes, the definition of evaluation perspectives, simulation tools and decision processes.

2.1 Definition of evaluation components

The basic geographical unit of the model is the land use parcel. Each of the components defined is classified according to the type of land use. This classification was based on the following evaluation components:

- Effects (E)- The actions resulting from human activity which are susceptible of decreasing the environmental quality of the studied area. In this project the following effects were considered relevant: Water release, Habitat destruction, Solid residue release, Noise emission, Air emission and Erosion;
- Attenuation scenarios (A) - An attenuation scenario represents the attenuating potential of each land use when related with one type of effect; One land use parcel may either attenuate or magnify specific effects that happen with its boundaries: Attenuation is represented by a value inferior to one; Magnification should be superior to one but it is not being considered currently. These values are also determined expertly, they are qualitative parameters and not spatial characteristics of propagation;
- Sensitivities (S) - The environmental quality components were classified and weighed against the land use classification producing the concept of Environmental Sensitivity (in this case Biodiversity, Quality of superficial water, Air quality, Soil quality, Acoustic quality and Landscape quality).

The sensitivities, effects and attenuation scenarios are evaluated, at the three user levels defined above, through a system of weights that qualifies them for each land use. This system allows the definition of the importance given to land use types. Each component's evaluation on the land use is translated into a values map (Fig. 1), a spatial classification of the existing land uses according to the evaluation perspective applied.

The definition of evaluation components is structured in evaluation profiles which can be interactively defined and modified. The evaluation perspective of one user can be stored and compared against others, enabling the experimentation and transparency capabilities of the model.

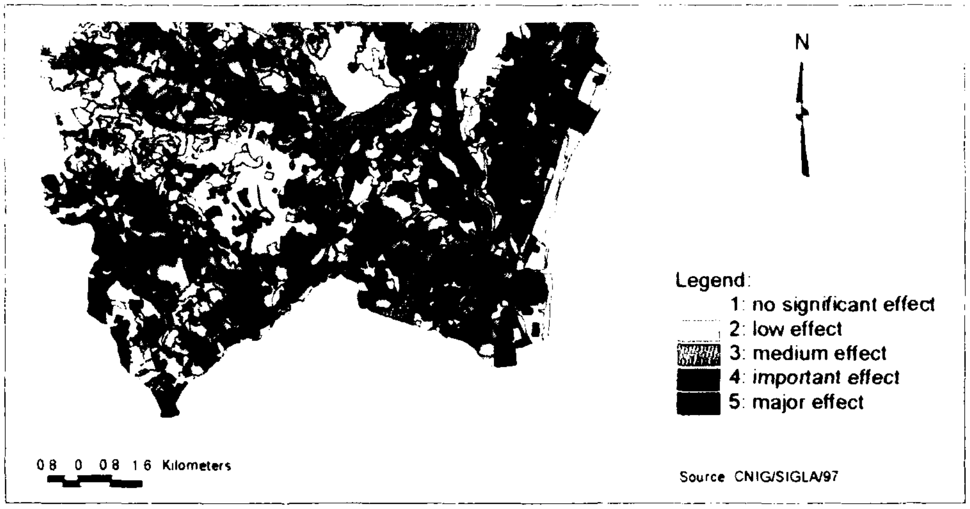


Fig. 1 Water Release effect values for the current land use map of the Southern area of the Loures Municipality

2.2 Simulation Tools

The simulation tools were development as a toolbox that enables the generation of different simulation lines and their integration according to a specific evaluation perspective.

2.2.1 Propagation Scenarios

A Propagation scenario (CP) is a map of the potential spatial diffusion of an effect resulting from a land use transformation. In this project, propagation can be effected through superficial water, air, underground water or landscape. The propagation scenarios are calculated from physical elements of space. The scenarios have been implemented using cartographic modelling processes.

2.2.2 Simulation lines

The components described above are combined to generate the intermediate and final results which are called Simulation lines (L_{ikjl}). The impact of one effect is calculated by combining the effect's value map with the associated sensitivity map, the chosen propagation scenario and the relevant attenuation values. The result is called a simulation line representing the potential environmental risk for the current set of land use parcels in one defined moment t.

The functional representation of the simulation can be expressed in the following way:

$$L_{ikjl}(LU_t) = S(LU_t)\ q\ E_k(LU_t)\ q\ CP(LU_t)\ q\ A_l(LU_t)$$

Where LU_t is the set of land use parcels representative of one moment t; q is the function that enables the combination of two simulation components (in this case grid multiplication). $S(LU_t)$, $E_k(LU_t)$ and $A_l(LU_t)$ represent the mapping of, respectively, one of the Sensitivities, Effects and Attenuation Scenarios associated with the current set of land use parcels; $CP(LU_t)$ is the representation of the chosen propagation scenario for this simulation. Although the number of possible simulation lines is extremely high, only the ones resulting from compatible components will be generated.

2.2.3 Simulation Integration

The definition of integration rules enables the estimation of a general situation or an oriented study, towards one or several of the defined components. It is possible to evaluate the results from simulations based on one specific theme or on a combination of themes. For example, the impact of a transformed land use parcel can be studied for all of the environmental quality components or oriented

towards one of them (biodiversity, etc). It is also possible to integrate similar simulations created according to different evaluation perspectives. This will generate solutions representing areas of agreement among different users.

2.3 The Decision Process

One of the main objectives of this project was the possibility of improving decision-making through the use of simulation tools, describing processes and discriminating options resulting in extensive forms of visualization, according to evaluation perspectives defined in a municipal planning process. In this section we will describe the decision tools which were conceived and developed using the simulation and evaluation modules of the system.

2.3.1 Environmental Risk and Performance

The potential environmental risk of the area results from the generation of simulation lines and enables the assessment of the development of area by identifying major risks and priorities of development. A reference simulation line is calculated (for time t) using the registered pollution sources. Additional simulations will be derived from this reference. Inverting the values of environmental risk generates the evaluation of environmental performance.

2.3.2 Visualisation of a Land Use Transformation Impact

Environmental performance and risks are represented as a $2^{1/2}$ D metaphorical model to increase visual perception and to allow the caracterization of the impact properties (figure 2).

2.3.3 Decision Parameters

One modification of the geographical elements (land use parcels) between time t and t+1 generates two simulation lines. The impact of this change can be measured by the difference between the two lines and the characterization of the resulting shape. This shape produces parameters for decision making as variations in Area, Volume and Depth. These parameters describe the importance of the impact and enhance the understanding of its distribution.

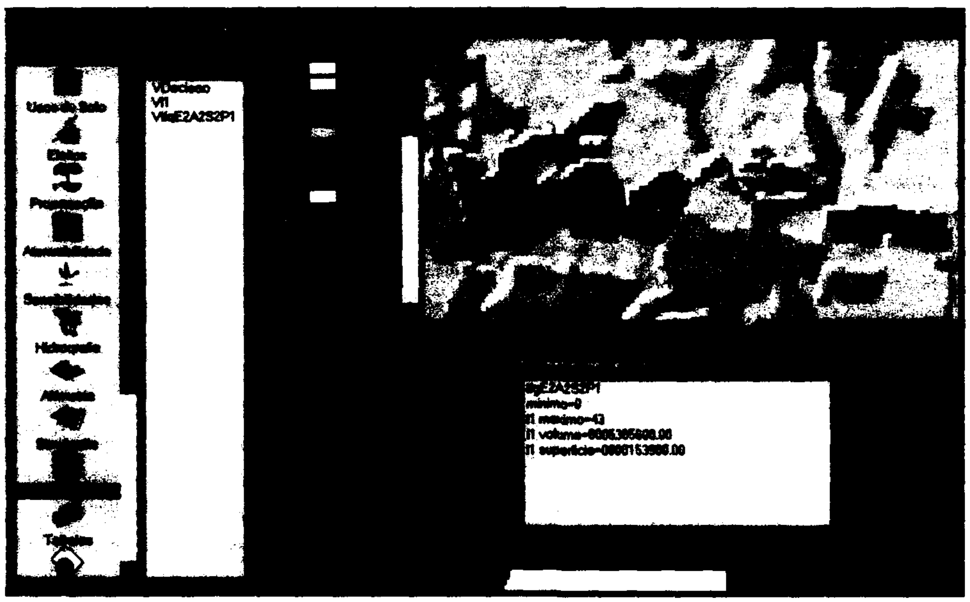


Fig. 2 - $2^{1/2}$ D representation of a simulation line $L_{a,t}(LU_j) = S_i(LU_j)\ q\ E_i(LU_j)\ q\ CP_i(LU_j)\ q\ A_i(LU_j)$
S = Sensitivity to Water Quality, E_i = Water Release effect, CP_i = Superficial water propagation scenario, A = Water Release Attenuation Scenario

2.3.4 Location problems

When concerned with a territory under study, the planner is often searching for the best place to locate a new plant, a new structure or trying to select priority parcels for remediation. The solution adopted in this project was to build a planning memory where simulation parameters and results are recorded. This memory is built from classifications of parcels provided by Land Use agents. These are intelligent agents which can evaluate their fitness for a specific land use change and bid for that change to be effected in their location. The fitness of each parcel/agent is ranked by its nature (type of land use), neighbourhood sensitivities, topological properties, and planning normative associated. Land use agents are currently under development.

2.4 System Architecture

The system control relies on a multi-agent system being built using Java and an associated Intelligent Agent library. Being a portable, object-oriented language, Java was an obvious choice for the development of the system, enabling the creation of modules than can easily be extended and dynamically changed. The intelligent agent library includes communication and reasoning as basic mechanisms allowing the developer to easily create and manipulate agents while concentrating on their behavioural characteristics. This architecture includes the modelling system, data storage and analysis tools.

2.4.1 The Dynamic Structure of the system

A multi-agent system structure was created to enable the system with dynamic and transitive connections between the components. When one spatial component or criteria changes during execution, this change will be reflected in all the components that depend on the former. Therefore, all these components must be updated. This operation is activated autonomously by the spatial agent responsible for the changed component. The system of dependencies is provided by a knowledge base of connection rules which is also updateable. Fig. 3 shows the structure of the system.

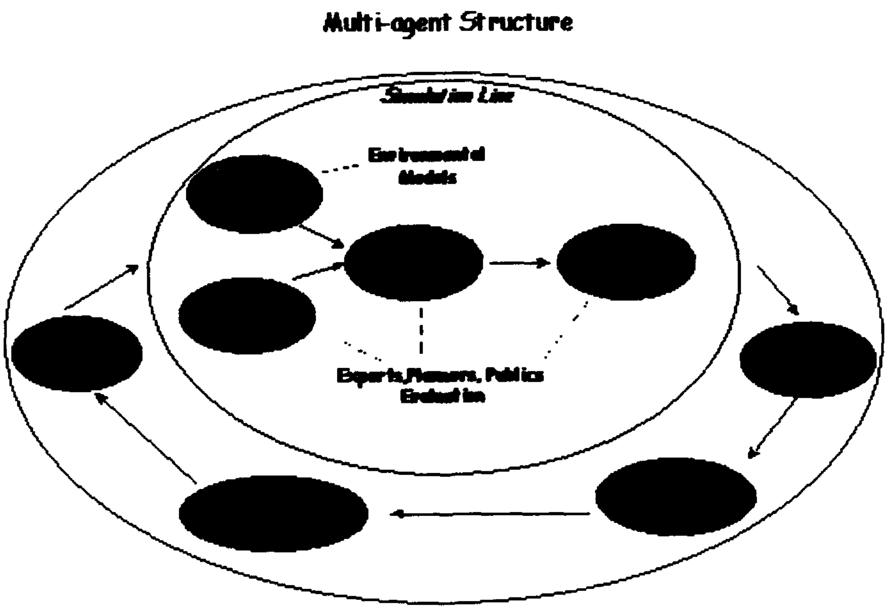


Fig. 3 — Functional structure of the system

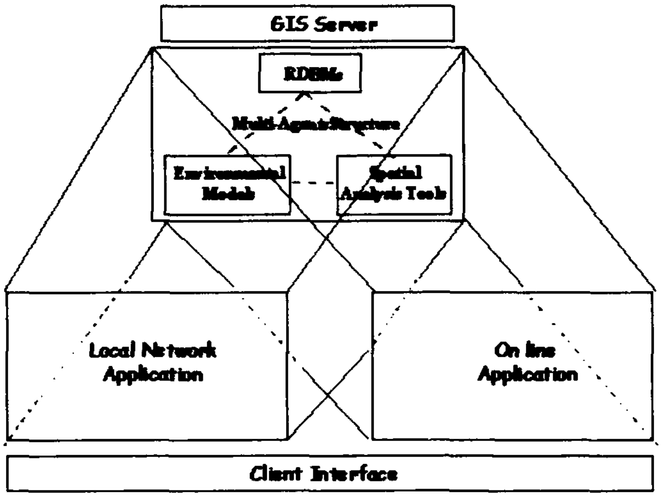


Fig. 4 - Application Design

2.4.2 Client Interfaces

The current prototype implementation has been built using the client-server paradigm . It has been constructed around a GIS server remotely accessed by two kinds of interfaces as shown in Fig. 4.

The local network application has been designed to work at the municipal level and enable the complete toolbox. An online application is also being created to store public proposals. It is an online mapping application which interacts with the land use data using evaluation profiles defined by the current user. The Java-based interface allows for the definition of the user's profile, the execution of simple simulations and the presentation of mapping results. This tool not only realises the transparent property but also constitutes a way to inform the public about the methodology used.

3 THE NEED FOR NEW DATA REPRESENTATION MODELS TO REACH HIGHER SIMULATION DIMENSIONS

Currently, the existing implementation does not solve all of the planning problems involved. The system implemented does not allow the cumulative impact evaluation of simultaneous land use transformations (Schweigert, 1994). Also, there are some problems with associating spatial and non spatial information. Finally, we have identified serious limitations in modelling non-continuous phenomena and transport mechanisms. The use of vectorial structures for modelling municipal information has clarified a need for new forms of representation that, not only handle vectorial information, but that can also enable non-contiguous forms of propagation. This will allow for the modelling of the complex environmental interactions involved as well as their temporal characteristics.

3.1 Heuristic Definitions

The definition of the data model is now underway and it will include the representations to be used and the properties associated with each object or class of land use (spatial and non-spatial elements of the system). A major effort is also being made to define a behavioural model for the land use object when confronted with a negative effect emitted by another object. These two models represent heuristic definitions from which the rules to be computed are explicitly created.

3.2 Data Modelling Issues

The data model required has to integrate topological de-

scriptions that will enable effective qualitative spatial reasoning (like distance and orientation description) related with the currently studied phenomena. The Object Oriented model will then appear as an appropriate structure to represent environmental interactions.

3.2.1 Object Orientation

The Object-Oriented (OO) model can be seen not only as an elegant alternative to the relational model but also as a solution for a more appropriate description of phenomena. The concept of object emerges from the necessity of manipulating, not only the static structures of information (data oriented) but also the dynamic behaviour of the system. Just like in the Entity-Relationship diagram, the static aspect of an object is presented as a collection of attributes. The set of attributes of one object is called its state. The dynamic and behavioural side of the object is presented through a set of operations (called methods) that will be executed under certain conditions. This possibility led the project to an OO approach, as the necessity for the definition of behaviour for different types of object became clear.

3.2.2 Minimisation of impacts

Hazard zones can also be represented as objects and their spatial interaction can be studied through topological and distance properties between different impact shapes and pollution sources. The objective is now to define rules and methods to minimize impacts, by reasoning spatially on the qualitative properties of the impact shape. Those methods will highlight the action to be taken at the pollution source to reduce the impact, measuring the simulation results through the decision parameters.

3.2.3 Temporal Aspects

Temporal representation is often limited in most of the systems used. However, new techniques have been recently proposed. In this project, we are considering the existence of different temporal versions of a same object generated by events (Wachwicz and Healey, 1994). Each impact is represented as a geometric object able to be represented by a different temporal version. Another interesting approach results from Peuquet and Duan's (1995) who consider an event-based spatio-temporal data model to handle forestry change. This approach is currently under a comparative study with the one suggested above.

3.2.4 Computational Implementation

The object characteristics of the model offer a few computational advantages. The most prominent one resides on the natural implementation of software interfaces and their flexibility. Moreover, the coupling process models facilitate the agents manipulation of the model elements. Finally, spatial, attribute and thematic relationships can be easily described and maintained by the feature itself as the input of new data (Crosbie, 1996).

4 CONCLUSIONS

It is our belief that the first implementation of the Decision Support System may lead to the reduction of the gap between the modeler's perspective and the decision-maker's habits. The opportunity to simulate changes in the studied area and experiment with new planning methodologies enables the planner with information and tools for simulating possible scenarios while the modeler can test new tools and receive feedback from the decision-maker. The main advantage is to get both parties to communicate and reach solutions together. The agent-based architecture can also act as a guiding mechanism, helping the user avoid evaluation parameters that cannot be applied in the specific problem (e.g.: avoiding decisions that go against normative). The evaluation perspective is not fixed and can interactively evolve in time. The planning decision process differs between countries but the difficulty of introducing new support tools is still the same in most contexts. This implementation completes the fundamental objective which was to propose a system that would integrate environmental matters in the planning process in a clear and scientific way. The implementation has been based on two components: a GIS server and a distributed access system. The GIS server uses a Multi-agent framework to link the modelling support system with the data storage and the

analysis tools. The Internet application allows the storage of public opinion and its familiarisation with such tools in a simple, inductive and educating way. Investment is now turned to the system's improvement through the exploration of new data structures and new process models. Another idea to be implemented is concerned with the calculus of the action to be taken at the source to reduce the studied impact in a qualitative way. We hope that future results will support us in the idea that new data models have to be imagined to increase the cognitive representation of space and our understanding of the Earth's mechanisms.

5 *ACKNOWLEDGMENTS*

The research performed by Armanda Rodrigues is funded by grant PRAXIS XXI/BD/2920/94 of the PRAXIS XXI/ JNICT Programme.

The SIGLA Project was funded by JNICT.

6 *REFERENCES*

Crosbie, P. (1996), Object-Oriented Design of GIS: A New Approach to Environmental Modeling, in *GIS and Environmental Modeling: Progress and Research Issues*, Carol Johnston, David Maidment, Michael Crane, and Sandi Glendinning, Editors, 1996, pp 383-386.

Janssen, R. (1991), *Multiobjective decision support for environmental problems*, Urije Universiteit te Amsterdam, 1991.

Peuquet, D., Duan, N. (1995), An event-based spatio-temporal data model (ESTDM) for temporal analysis of geographical data, in *International Journal of GIS*, 9 (1), 1995, pp 7-24.

Rodrigues, A., Grueau, C., Raper, J., Neves, N. (1997), Research on spatial agents, *proceedings of the third joint European conference on geographical information*, Vienna, Austria 1997, IOS Press.

Schweigert, D.(1994), Multicriteria Aspects of Environmental Indices, *Applying Multiple Criteria Aid for Decision to Environmental Management*, Euro Courses: Environmental Management, Volume 3, Paruccini, M. (ed.), 1994.

Shiffer, M.J. (1992), Towards a Collaborative Planning System, in *Environment and Planning B: Planning and Design*, 1992, pp 709-722.

Wachowicz, M. and Healey, R.G. (1994); Towards temporality in GIS, in *Innovations in GIS*, ed. Worboys, M., Taylor and Francis, London, 1994, pp 105-115.

The GIS revolution, are we ready for GeoComputation?

Samuel Mann
Spatial Information Research Centre,
Department of Information Science, University of Otago, NZ
samann@commerce.otago.ac.nz

Presented at the second annual conference of GeoComputation '97 & SIRC '97, University of Otago, New Zealand, 26-29 August 1997

Abstract

This paper examines the argument that the GIS revolution has run its course. An approach is used that focuses on describing requirements rather than finding new roles for increased processing power. Results of a survey of resource management practitioners are described. A prototype system, SPMS is also described which was developed from a position of information requirements. The system is shown to be useful. The implications of this work for developers of Geocomputation, both direct and conceptual, are discussed. This should result in useful development of new tools in Geocomputation.

Introduction

The author is a keen supporter of increased use of supercomputing (or High Performance Computing; HPC) in geographical research and practice. This paper stems not from fear, but a desire to see that such advances are appropriately delivered to the desks of geographical practitioners. The flyer for this conference argued that "the GIS revolution has run its course", that "there are few tools around today that are of any real use" and "the problem is that most researchers in GIS have largely overlooked the impact of new technologies based upon high performance computing". This paper takes exception to this argument and suggests an approach that sees the development of useful tools that may or may not include supercomputing.

This paper presents results from a case where GIS is poorly used despite high computer competence and GIS availability. The data shows that for environmental managers it is not a speed of processing problem but one of the computer, GIS in particular, missing the point in what the practitioners are trying to achieve. This paper will show that a redesign of GIS in line with practitioners problems results in tools "of real use". This is in keeping with Landauer's (1996 p131) argument that researchers and developers have fallen into the habit of saying "here is something I can do with my computer, I'll do it". Landauer argued that the approach should be "here is something we wish we could do, but gosh, how can you do it?".

This paper is in two sections, the first examines the argument that the GIS revolution has run its course. The second examines what planners of new research and solutions can take from efforts to improve the state of current GIS. This includes the 'new challenges' posed by existing problems.

GIS use and adoption

To argue that the GIS revolution has run its course is to overlook a crucial part of a revolution, its acceptance by the people. While numerous surveys might show the widespread adoption of GIS and some researchers may think they have all the problems solved, the few surveys on actual GIS use suggest that the diffusion is far from complete.

In defining "Geocomputation"[1] Openshaw and Abrahart (1996) argued to the effect that GIS technology has had

[1]Though note that they never actually managed a definition.

its day. They placed great emphasis on the fact that computing speeds are now 10^6 times greater now than during the GIS revolution. It was implied that GIS has met its objectives and we (the researchers) should now move onto something else. This is examined with respect to the use of GIS for environmental management, particularly regional environmental decision-making (REDM).

GIS is widely promoted as a suitable tool for environmental assessment and analysis. This appears to be backed by the finding of Marr and Benwell (1996) that 70% of New Zealand local government authorities have GIS. Further, previous surveys have pointed to a wide availability of environmental data (Benwell and Mann 1995) and that this data is increasingly considered 'mature' (Marr and Benwell 1996). Their figures, however, also included some disturbing trends: while 45.5% of authorities intended using GIS for Land Use Mapping, only 20% were actually doing this, a significantly greater fall than for other intended uses (Mann 1997). This trend is not restricted to New Zealand, Campbell (1994), found a lack of analytical capability in British local government GIS, most being used only to support basic display and query. Zwart (1992) also found that of 142 systems available, 70% were only capable of display in mapped form the results of textual rather than spatial manipulations.

The most recent survey in this area (Mann 1997) aimed to examine computer support for environmental decision-making, particularly the obstacles for such use. The subjects were New Zealand resource management practitioners. The respondents (n=82) were shown to have a high level of computer use (97.3%) and use a wide variety of systems and software. The desktop office was dominant with 82.3% of respondents using word processing as a 'primary use'. Spreadsheets, electronic mail, database use and drawing scored between 52% and 32%. GIS was considered a primary use by 12% of respondents. More than two-thirds of the problems people were working on took more than a week to complete, many took considerably longer. There is, however, a large level of dissatisfaction with computer support for their work, with 35.3% dissatisfaction of the total sample rising to 46% of an identified group of Environmental Managers (EM).

When asked to rank their computer functions in order of major uses (from a given list), 23% identified GIS. When asked later in the questionnaire whether they used GIS in their work, a higher number, 45.59%, indicated that they did. This difference in results suggests that people do not yet consider GIS to be part of their desktop as suggested by Somers (1996). Further, when asked to what use they put GIS, an extra group indicated that 'they didn't actually use it but that somebody else did for them'. This is described by Somers as 'chauffeur-driven GIS' and is generally considered to be an undesirable situation. Campbell (1994) pointed out the disadvantages in this approach as "technical specialists generally [have] little understanding of the nature of information required by users, simply perceiving it to be units of data" (p 319).

Respondents agreed with a number of statements concerning GIS: that they knew what is understood by GIS, that is used in their organisation, that it is essential for their work, and that their organisations were generally supportive of GIS use. So what are the obstacles to the successful use of GIS? A number of options are apparent, only one of which is a need for increased computing power.

Respondents were asked to choose from a list of potential advances in order to complete the phrase, 'My work would be made easier by the development of software with...'. Although no single advance was desired by a majority of respondents, 'Friendly interfaces' and ' Focus on environmental processes' were at the top of the list with 48% and 45% respectively (Table 1). The computing technical problems; data structure, speed of processing and faster graphics were at the middle or bottom of the list. Cost was not a prime consideration, and perhaps surprisingly, neither was artificial intelligence. The preferences of the EM group differs from the total sample. The main effect is the demotion of 'better map production capabilities'. The three most desired improvements for this group are for a change to 'focus on environmental processes rather than computer operations', 'friendly interfaces' and

Friendly interfaces (48)	Focus on env process (54)
Focus on env process (45)	Friendly interfaces (49)
Map production (40)	Helps communicate decisions (40)
Helps communicate decisions (31)	Approaches represent problems (34)
Uncertainties in data (31)	Knowledge bases (31)
Knowledge bases (30)	Confidence of result (31)
Easier to learn (28)	Map production (29)
Efficient data structures (25)	Easier to learn (29)
Faster processing (25)	Easy methods of input (29)
Approaches represent problems (25)	Uncertainties in data (26)
Easy methods of input (25)	Efficient data structures (20)
Confidence of result (24)	Faster processing (20)
Cheaper (22)	Non spatial (20)
Faster graphics (19)	Cheaper (17)
Non spatial (13)	Faster graphics (11)
Artificial intelligence (10)	Artificial intelligence (11)

Table 1 Software preferences for total sample and Environmental Managers

software that 'helps communicate how decisions are made'. Again, the engineering improvements scored lowly.

The total sample ranked 'with more friendly interfaces' highest in their software preferences. For environmental managers, it ranked second behind 'focus on environmental processes'. Nielson (1993) points out that a system "does not have to be friendly, just not get in the way of their work" (p 23). In terms of this not getting in the way of work, the high ranking of the two software preferences, 'with a focus on environmental processes rather than computer operations' and 'with approaches that better represent the problems I work on' is somewhat removed from the argument that we must adapt our ideas to parallel processing.

Improved GIS

The findings briefly presented above were used by Mann (1997) as part of a project that aimed to improve the support of regional environmental decision-making. This project successfully took the approach of placing much effort on identifying needs before specifying tools. The resulting tool is briefly presented with a view to examining what implications emerge from this work for the broader area of GeoComputation.

After examining current decision-making, a decision building environment was proposed and a set of conceptual criteria were derived that, it was argued, would lead to improvements in decision-making. Existing support systems were examined and it was shown that a major obstacle is the lack of a system that integrates components of GIS with process modelling functions, particularly those used in Visual Interactive Modelling Systems (VIMS, Pidd 1996). A new approach, Spatial Process Modelling was proposed that combines GIS with VIMS. Detailed design specifications were developed that were used to develop a prototype system. This system, the Spatial Process Modelling System (SPMS) allows users to build complex environmental models simply by drawing diagrams (see Figure 1). These diagrams consist of three components, spatial objects (maps), data objects and process objects. These components are linked together to form the model structure. The system interprets the diagram and performs the processing, after processing the thumbnail maps are updated with new values were needed. As models may in-

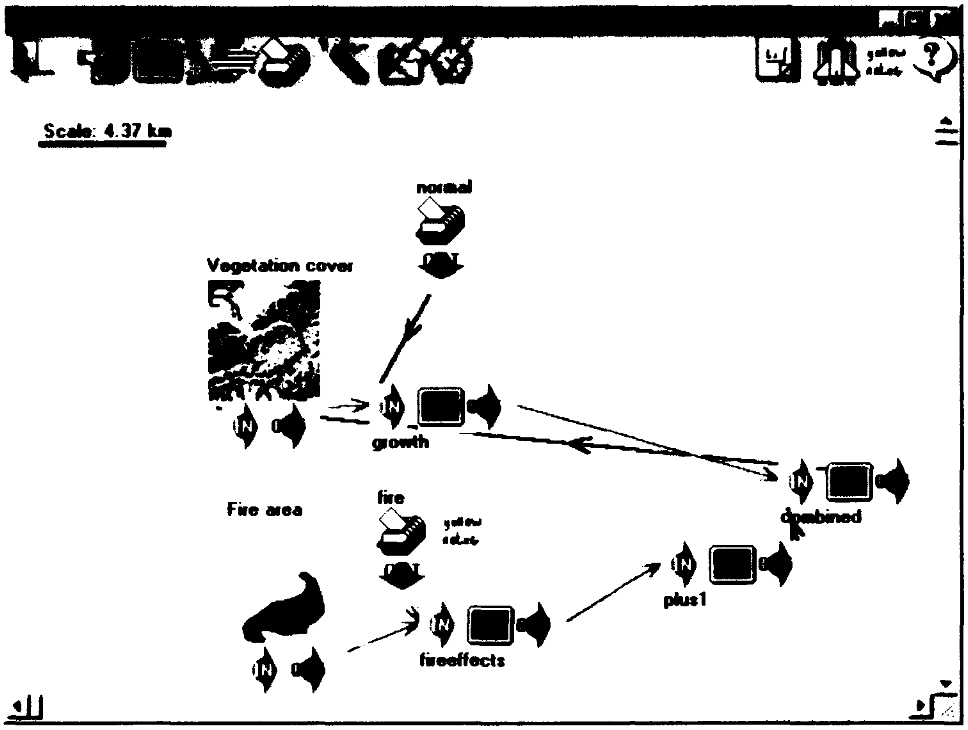


Figure 1 Spatial Process Modelling System with a user's fire response model

clude feedback loops and be defined for any time period, scenario development and prediction testing is possible (cf. systems such as Albrecht *et al.* (1997) which do not allow feedback and are therefore workflow representations). It is intended that the SPMS be used by individuals or in workshops.

In Figure 1 the user's model shown represents the effect of burning on vegetation growth. The user first built a simple model to predict vegetation growth. The vegetation map at the top left is joined to a growth component along with a data file icon which represents growth conditions over 15 years. This user's model has a time step of a year, and is set to run for 15 cycles (years). Originally the user had feedback from the growth component back into the vegetation map which was updated for each cycle. To explore the effect of burning a part of the landscape, the model was adapted to include a burning component. The lower map on Figure 1 is an area to be burnt. The fire data icon represents a hypothesised burn response curve (over the 15 years), because it is hypothetical and may be controversial it is annotated with appropriate comments. The response curve is combined with the affected area, returned to a multiplier and combined with the normal growth before being fed back into the vegetation map. This model may be run and shows that the vegetation does not fully recover over the 15 years. The user may then explore the effects of changing the fire response curve or growth conditions, or adapt the model structure to investigate, say, the effects of including altitude in the system.

In order to test whether the proposed decision building environment embodied by the prototype SPMS resulted in measurable benefits in decision-making, the SPMS was evaluated for its contribution to decision-making standards. Practitioners in environmental management used the SPMS to complete some model building tasks taken from current issues in environmental management. They were

then asked how the system would function in terms of decision-making standards. They found that the system would help in understanding the problem, developing solution criteria, using relevant information, generating a range of alternatives, exploring uncertainties and in examining consequences.

The perceived potential benefits of the SPMS may be related to a number of factors. The ease of use and overall satisfaction suggest that the users' model is matched by the system model (Pidd 1996). It is also a definite move in the direction of Davies and Medyckj-Scott's (1994) aim to break down differentiation in terms of control and user display representations. Further, in Woodmansee's (1988) terms, the user can not only envision several layers simultaneously, but also the links between layers are made explicit.

Participants were asked whether the system would be a useful tool in described environmental management scenarios. The scenarios provided a check that the experimental tasks, which had centred on modelling, were seen to fit into the wider aspects of decision-making. One scenario was adapted from an issue with which some participants had current involvement. It was stated 'In preparing a regional plan, a planner considers the effect on river flow of potential forestry plantings in medium sized catchments with summer low flows. Knowing that trees reduce runoff, it is proposed that plantings be limited to below 300m'. All participants responded that using the SPMS in this scenario would be useful. A policy analyst, for whom this was a real and current problem, commented "That's exactly what the problem needs". This differs slightly from the responses of others who saw this scenario more from an analytical perspective; "may be a complicated model with considerable scientific data". Nevertheless, they all still thought it feasible and beneficial. The differences reflect a difference between the group who see the model results as important and those who see the modelling itself as important.

When participants were encouraged to consider features for future development of the SPMS, they responded that while they could see the benefits of altering components to test model structure and the annotation of this action, they felt this procedure should be formalised in some way, commenting; "How would you find out what you left out?" and "Examining consequences dependent on data quality to assist in analysis of uncertainty".

Another area of discussion was over a conflict in perception, is the SPMS a tool for detailed analysis or one for exploration and discussion? It is believed it can be useful for both, an assertion backed by one participant's comment; "this is a big step in getting both discussion and [analytical] quantification beyond hearsay and straight bickering". If the SPMS is to be used for analysis, methods for the validation of structure and investigation of error are needed. The current system offers little beyond Rothenberg's (1991) 'naive perturbation' and the flexibility to change model structure and dynamics. Automated sensitivity testing, whereby the value of each variable is systematically perturbed and the effects on results examined, would be beneficial. This may, however, be prohibitively time consuming. One option would be to develop an architecture whereby, once a model is constructed, it is sent off to a more powerful computer for immediate or delayed batch processing. This would require a protocol for the transfer of models and data such as that currently under development by Marr *et al.* (Geocomputation '97).

The validation of developed models is unlikely to be a completely objective process. Pandey and Hardaker (1995 p 446) pointed out that as "no model can be like the real system in the sense of being identical with it...the problem therefore is to decide whether a particular model is 'good enough'". As 'good enough' means that it performs "acceptably closely to the real system in some selected, important respects" and the selection of these important aspects relates back to the purpose of the modelling, the model validation itself is a "subjective and uncertain process". How to support this validation process without interfering with benefits of the simple VIMS approach should be a focus of future research.

One of the questions posed in the development of the

specifications for the SPMS concerned how much complexity should the user be presented. The participants in the study found the complexity appropriate. The problem of the need for mathematical work-arounds (for example, the 'plus 1' component in Figure 1) would be reduced by the 'equation builder' including more complex equations. A new release of Idrisi (Eastman 1997) has an 'image calculator' which allows the construction of extended equations for "creating and running GIS models". It contains only calculator functions (i.e. non spatial) and does not permit feedback loops, both factors meaning complex spatial models cannot be constructed, but it is a step in the right direction.

The decision building environment was intended to assist with what Lowes and Walker (1995) called "novel problems" that Fedra and Reitsma (1990) described as not justifying unique solutions. The SPMS may be considered a simple programming language capable of rapid construction of flexible models. If the SPMS was to be used regularly in a region dealing with similar or related issues, it may be beneficial to develop a library system for larger pre-built model components. This might include components representing, for example, a tussock growth model, a sheep model or river flow model. It would be a fine line though, these components should be able to be broken down and have assumptions changed and not become large black box components. A further, complementary option would be to employ templates for models. Such templates might give the structure and dynamics of generic 'river flow' or 'grazing' model.

One of the important advances this study has shown is the benefits of incorporating time as an integral component in spatial modelling. This has allowed the construction of models that include feedback and prediction over a number of cycles. These cycles may be named 'years' or 'months' but this is arbitrary. The system has no inherent 'knowledge' about the relationship between a day and year. There is a need to further develop the temporal concepts to allow more complex arrangements of time/class data objects (to support, for example, grazing management charts) while retaining the iconic VIMS approach. One way to proceed with this may be to allow the user to embed models with different time-steps. This may also allow the development of non-linear models.

The emphasis on modelling in a visual environment facilitates the expression and exploration of understanding but it still comes down to a matter of defining mathematical relationships (interactions) between objects. The SPMS achieves an ability to represent model structure in a way that better represents methods of environmental management but still forces a numerical approach. This may not appropriate for all scenarios. It would be difficult to represent a decision that is expressed in the form "I move the ewe flock when I can see the tussock flowers from the woolshed". It is worth noting, however, that while this may be difficult, it would not be impossible. Influence diagrams, which operate on 'this factor has a positive/negative influence on this factor' remove the need for mathematical expression but accordingly, are not operable simulation models. Other options include natural language processing, graphs drawn by hand or other less mathematically less rigid methods. Perhaps a system that operates on different levels would allow the appropriate representation for each case.

The explicit representation of model structure goes some way in reducing uncertainty. The representation of structure and tools for annotation used by the SPMS provide a reduction of the problem described by Robertson *et al.* (1991 p 1) as "current modelling tools do not provide adequate support for documenting important modelling decisions and this makes it difficult for them to be understood by others". The appropriateness of decisions though still relies on the REDM practitioners (and stakeholders) judging the quality and relevance of description of the problem (aka the model). Assistance may be incorporated directly into computer systems or be used by a facilitator. This might include decision-making criteria based on normative standards; 'have you included all available information?' or model quality criteria such as those discussed by Openshaw (1995).

If such context based assistance (Hoschka 1996) was em-

bedded in a system, it could be amalgamated with components that facilitate the development of models. Walters (1986) and Grayson *et al.* (1994) described the process by which models may be developed. They suggest identifying key indicators and then articulating variables and processes that directly affect the indicators and continuing until boundaries are reached where it is felt further elaboration would unnecessary. While the SPMS encourages the development of a model following this approach it is not actively facilitated. A similar approach to model development is to define high level relationships; in a pastoral system model example this might be a grazing/growth relationship. Each component may then be further defined or 'exploded'. While this is a simple concept on paper or a whiteboard, such hierarchical processing is a complex computation task but has been successfully applied to Petri nets (eg. Purvis *et al.* 1995).

Lessons for Geocomputation

There are a number of implications of this work for developers of Geocomputational approaches. The first are direct implications, while a second group acts on a more fundamental level.

Many of the limitations of the current SPMS are related to an ability to process spatial data more effectively. This is particularly the case for sensitivity testing. A grand challenge, then, would be to develop an architecture whereby systems such as the SPMS could make use of increased resources.

Maxwell and Costanza (1995) used parallel computers to drive models generated in the VIMS, Stella, for every cell in a landscape. Bridging programs allowed communication between cells such as for the lateral movement of water. This, however, is a large scale solution, and beyond the capabilities of REDM practitioners. Westervelt *et al.* (1995) describe a similar Dynamic Spatial Ecological Modelling (DSEM) system. They report that while model development within the VIMS (again, Stella) facilitated collaboration and model design, those parts of the process that required a (FORTRAN) programmer, meant "it is all to easy to lose track of what the program is actually doing" and "discourages efficient changes and modifications".

Mineter and Dowers (1996) discussed a layered approach for parallel processing where "parallel libraries in effect hide the parallelism from the developer of an application, and so reduce the parallel computing expertise demanded of that developer" (p 602). If such a layer was to be incorporated into a buffer computer in much the same way as for distributed database design, jobs could be submitted over the internet from systems such as the SPMS in instances where the processing may otherwise become too strenuous for local processors. This would allow a shift in emphasis and a redefinition of 'users' and 'developers'. The SPMS has shown that practitioners (users), can develop models that would have previously been the domain of programmers (developers). Effort in Geocomputation should be aimed at facilitating this transfer in HPC.

Landauer (1996 p 6) described phase one of computing whereby "computers can do anything that can be reduced to numerical or logical operations" and that these 'easily reached fruits have been picked". "Helping people think" is the next goal and is reflected in the conceptual criteria as 'Emphasis on facilitating human interaction and thinking for both workshop situation and single user'. Burrough and Frank (1995 p 105) argued that "there is a large gap between the richness of the ways in which people can perceive and model spatial and temporal phenomena and the conceptual foundations of most commercial geographical information systems". This is only worsened by expecting users to port their understanding to high performance computers (e.g. Turton and Openshaw 1996).

Conclusion

This paper has argued that the GIS revolution is not over. A greater use of HPC in geography is supported but progress does not, however, rely on a few enthusiasts porting existing models to more powerful computers while the real users become frustrated with inappropriate conceptions of their problem space. An example was described of a GIS modelling system built to problem solving crite-

ria. It allows predictive and exploratory modelling of environmental systems in a manner that is not fast but allows something to be done that could not be achieved before. In its present configuration, the system lacks sensitivity testing as this is indeed prohibitively time and resource consuming. An architecture is presented that makes use of the internet to remotely use parallel computers where appropriate as part of the normal workspace. Supercomputing and GIS can become closer, even to the extent of forming a new field, but this 'GeoComputation' should be seen as part of a move to empower users not just to give them more power.

References

The prototype Spatial Process Modelling System software described *in this paper may be obtained via the internet* from:

http://divcom.otago.ac.nz:800/sirc/spm/spms.html

Albrecht, J. H., Jung, S. and Mann, S. (1997). VGIS: a GIS shell for the conceptual design of environmental models, (in) Z. Kemp (Ed) Innovations in GIS 4, Taylor and Francis, London, p154-165

Benwell, G.L. and Mann, S. (1995) Advances in Spatial Information Technology & Implications for the Telecommunications Industry, Spatial Information Research Centre, University of Otago, Dunedin, New Zealand, 125p

Burrough, P.A. and Frank, A. U. (1995). "Concepts and paradigms in spatial information: are current geographical information systems truly generic." International Journal of Geographic Information Systems 9(2): 101-116

Campbell, H. (1994). "How effective are GIS in practice? A case study of British local government." International Journal of Geographical Information Systems 8(3): 309-325

Davies, C. and Medyckyj-Scott, D. (1994). "GIS usability: recommendations based on the user's view." International Journal of Geographical Information Systems 8(2): 175-189

Eastman, J. R. (1997). IDRISI for windows (promotional material for version 2). Clark University Graduate School of Geography, USA, http://www.idrisi.clarku.edu

Fedra, K. and Reitsma, R. F. (1990). Decision Support and Geographical Information Systems, (in) H. J. Scholten and J. C. H. Stillwell (Eds). Geographical Information Systems for Urban and Regional Planning, Kluwer, Dordrecht, p177-188

Grayson, R. B., Doolan, J. M. and Blake, T. (1994). "Applications of AEAM (Adaptive Environmental and Management) to Water Quality in the Latrobe River Catchment." Journal of Environmental Management 41: 245-258

Hoschka, P., Ed. (1996). Computers as assistants: a new generation of support systems. Computers, Cognition, and Work. New Jersey, Lawrence Erlbaum Associates. p313

Landauer, T. K. (1996). The trouble with computers: usefulness, usability and productivity. Bradford MIT, Cambridge Massachusetts, 425p

Lowes, D. and Walker, D. (1995). Environmental Management: Opportunities and Challenges in the Application of AI. Proceedings of the Workshop on AI and the Environment: Eighth Australasian Joint Conference on Artificial Intelligence, C. McDonald (Ed) Australian Defence Force Academy, Canberra, University College, University of New South Wales. 1-11

Mann, S. (1997). Spatial Process Modelling for Regional Environmental Decision-Making. Information Science. Unpublished PhD Thesis, University of Otago, Dunedin, New Zealand 269 p http://divcom.otago.ac.nz:800/sirc/spm/spms.html

Marr, A. J. and Benwell, G. L. (1996). "The evolution of the use of GIS in New Zealand Local Government." New Zealand Surveyor (286): 30-35

Marr, A. J., Pascoe, R. T. and Benwell, G. L. (1997). Interoperable GIS and Spatial Process Modelling. Geocomputation '97, R. T. Pascoe (Ed), University of Otago, Dunedin, New Zealand.

Maxwell, T. and Costanza , R. (1995). "Distributed Modular Spatial Ecosystem Modelling." International Journal of Computer Simulation: Special Issue on Advanced Simulation Methodologies 5(3): 247-262

Mineter, M J. and Dowers, S. (1996). Parallel processing for geographical applications: a layered approach. 1st International Conference on Geocomputation, R. J.Abrahart (Ed) University of Leeds, Vol 1 601-612

Nielson, J. (1993). Usability Engineering. AP Professional, Cambridge, MA, 358p

Openshaw, S. (1995). "Human systems modelling as a new grand challenge area in science (Commentary)." Environment and Planning A 27: 159-164

Openshaw, S. and Abrahart, R. J. (1996). Geocomputation. 1st International Conference on Geocomputation, R. J. Abrahart (Ed) University of Leeds, 665-666

Pandey, S. and Hardaker, J. B. (1995). "The role of modelling in the quest for sustainable farming systems." Agricultural Systems 47: 439-450

Pidd, M. (1996). Model development and HCI. 1996 Winter Simulation Conference, Hotel Del Coronado, Coronado, California, The Society for Computer Simulation, Institute of Industrial Engineers, The Institute of Electrical and Electronics Engineers. 6 pages

Purvis, M. K., Purvis, M.A., Benwell, G. L. and Mann, S. (1995). Modelling and Simulation of a New Zealand Legislative Act. International Congress on Modelling and Simulation, P. Binning, H. Bridgeman and B. Williams (Eds), University of Newcastle NSW, 4: 313-318

Robertson, D., Bundy, A., Muetzelfeldt, R., Haggith, M. and Uschold, M. (1991). Eco-logic: logic based approaches to ecological modelling. MIT Press, Cambridge MA, 243p

Rothenberg, J. (1991). Knowledge-based simulation at the RAND corporation. (in) P. A. Fishwick and R. B. Modjeski (Eds). Knowledge-based simulation - methodology and application. Springer-Verlag, New York, p133-161

Somers, R. (1996). GIS Management Strategies and Issues. SIRC'96: Proceedings of the 8th Annual Colloquium of the Spatial Information Research Centre, R. T. Pascoe, N. C. Sutherland and P. Gorman (Eds), University of Otago, 1-7

Turton, I. and Openshaw, S. (1996). Human systems modelling: results of the HPC initiative at Leeds. 1st International Conference on Geocomputation, R. J. Abrahart (Ed) University of Leeds, Vol 2: 828-830

Walters, C. (1986). Adaptive Management of Renewable Resources. Biological Resource Management MacMillan, New York, 374p

Westervelt, J. D., Hannon, B., *et al.* (1995). Dynamic, Spatial, Ecological Modeling. A Demonstrated Simulation of the Sage Grouse Habitat at the Yakima Training Center, Washington. (USACERL Technical Report 95/16): US Army Corps of Engineers' Construction Engineering Research Laboratories. http://blizzard. gis.uiuc.edu/htmldocs/GrouseModel/Grouse.html (Dec 1996)

Woodmansee, R. G. (1988). Ecosystem processes and global change. (in) T. Rosswall, R. G. Woodmansee and P. G. Risser (Eds). Scales and Global Change, Wiley and Sons, New York, p11-27

Zwart, P. (1992). Mature GIS - or taking the G out of GIS. Proc 4th Annual Colloquium of the Spatial Information Research Centre, G. L. Benwell and P. G. Firns (Eds), Dunedin, NZ, Otago University. 1-9

Predictive Assessment of Neural Network Classifiers For Applications In GIS

Gordon German, Mark Gahegan and Geoff West
Department of GIS, School Of Computing
Curtin University, P.O. Box U 1987, Perth 6001
Western Australia
email: gordon@cs.curtin.edu.au
Tel: +618 9266 3475, fax: +618 9266 2819

Presented at the second annual conference of GeoComputation '97 & SIRC '97, University of Otago, New Zealand, 26-29 August 1997

ABSTRACT

Artificial Neural Networks (ANNs) are well suited to implementing supervised classification tools for GIS data. They make no assumptions about the statistical nature of the data, can be used with ordinal and nominal data types together and can be trained with comparatively few training points, as they do not have to choose a data distribution model, unlike techniques such as Maximum Likelihood Classification.

However, training these neural network classifiers can be a time-consuming process, with no guarantee of the outcome. In this paper, the author presents a methodology for determining whether learning is practical for a given network on a given data-set, prior to commencement of the training phase. This is achieved by examining the error scores at the initial class boundaries and checking for redundancy in the network hyperplanes. This redundancy indicates how much flexibility is available in the network to learn complex boundaries.

1.0 Introduction

Neural networks have been used by the GIS community now for several years as an alternative tool for classification and feature extraction (Lees, 1994). Many commercial neural network software packages are currently used by the GIS professional and the next few years will likely see more of these classifiers integrated into GIS packages themselves. There now exists many variants of the original neural networks, as first proposed in the 1960s (eg. Amari, 1967), with a plethora of methodologies available to confuse the inexperienced user. This article is concerned with classification strategies and predicability of neural network classifiers, specifically in regard to our own neural net package, DONNET, which stands for Discrete Output Neural NET and is available via the World Wide Web on *http://www.curtin.cs/gis*.

Neural networks have often suffered from the dual spectres of difficulty-of-use and lengthy training times (Skidmore, 1995; Wray, 1996). In addition, the abundance of variations to the basic neural net paradigm available, makes it difficult for users not involved in the field to select the package most appropriate to their needs. DONNET is a neural network package specifically targeted towards users wishing to classify GIS data, across a range of statistical types (eg remote-sensed images, environmental surfaces, rock/soil classifications etc). In this paper, we examine how the learning phase associated with DONNET can be streamlined to the point where the GIS user can determine the suitability of the methodology to his/her particular task prior to actually starting the learning phase. The paper is organised as follows:

Section 2 describes DONNET's position in the hierarchy of artificial neural networks and gives a brief overview of the differences between DONNET and other networks. Although the final aim of the DONNET project is to produce a 'black box' classifier, it is often instructive to under-

stand the broader machinations of such a tool. Section 3 is a refresher on the basics of neural network operations as they pertain to DONNET, briefly presenting the functional and conceptual models. Section 4 introduces methods of rating classification. We then introduce a methodology for assessing the learning capability of DONNET as a GIS classifier prior to training, using the data analysis abilities inherent in the network to highlight problem areas. An extension to the model is considered in Section 5. Finally, Section 6 presents the conclusions and the direction that further work will take in this field.

2.0 The Neural Network Hierarchy

2.1 Application Types

Neural networks cover essentially two broad automation tasks in Artificial Intellegence - function approximation and pattern recognition (Pao, 1989). In the former, the task is to learn a specific function of arbitrarily many dependent and independent variables from a set of known relationships and then interpolate for unknown dependent variables. In the latter, the task is to learn to recognise several distinct generalisations of specifically presented patterns and then pick these learnt 'classes' from a set of unknown patterns. In terms of actual architecture and implementation, these two types of nets can be quite similar, (if not identical), but the way in which their functional characteristics are modelled is quite different. This conceptual modelling is, of course, merely a tool to help in our understanding of the relationship between the network's processes and the task at hand, so should not be considered as a literal description of the network's implementation. DONNET is firmly geared towards the pattern recognition task, although with some modification, it can be implemented as a function approximator.

2.2 Classifiers

Within the domain of pattern recognition, or classification, there exists two main families of classifiers, *supervised* and *unsupervised*. This parallels the hierarchy within the statistical classification scheme, where such methodologies as K-means clustering and Maximum Likelihood Classification (MLC) provide examples of unsupervised and supervised classifiers respectively. An unsupervised classifier is free to pick its own generalised pattern structures (classes), i.e. it separates the given data into classes based on similarities it distinguishes within certain groups of examples of the data. As such, the user does not have any control on how to classify the data, or how many classes the data will be partitioned into. A supervised classifier is provided with a 'teacher', normally in the form of a file of target classes for each member of the training set. In this way, the user can control how the data is to be partitioned and, for this reason, supervised classification is generally more common for the classification of GIS data.

DONNET is a multi-layered perceptron (MLP) configured as a supervised classifier. It is customised for handling large, poorly separable datasets, with small sample sizes, corresponding to limited ground truth.

3.0 DONNET - An Overview

3.1 Introduction

Most of the material in this section has been covered in more detail in previous papers by the author (German, 1995; German & Gahegan, 1996) and others (Bischof et. al., 1992; Dunne, 1993). Here we simply present the basic concepts of DONNET, as viewed from both the functional and conceptual models.

3.2 Functional Overview of the MLP

The architecture of DONNET consists of an input layer, a hidden layer and an output layer. The input layer is passive, whereas the nodes of the hidden and output layers perform sigmoidal transformations of their input data. The number of input layer nodes (*p*) represents the dimensionality of the feature space, whilst the number of nodes in the output layer (*q*), is the number of partitions, or classes, we wish to impose upon this feature space. The number of hidden layer nodes (*h*), is dependent on *q* as:

$$h = (q \times (q-1))/2 \qquad q>2$$

the formulation and reasoning for this is presented in German & Gahegan (1996) and Gahegan et. al. (1996). Hence

the architecture of the net is completely determined from the training data and the number of required classes, without assuming a particular distribution for the data.

Supervised classifiers learn by trying to minimise the error between the target set of outputs provided by the supervisor, and the learnt representation of the data by the classifier. DONNET's learning phase (as with most MLPs) can be analysed as four separate sub-phases:

1. For each sample (the input vector) in the training-set:
   a) Propagate the vector through the net.
   b) Compare the outputs with the target values and calculate an error figure.
   c) Back-propagate the error throughout the network's weight connections to give the *gradient* (derivative) of each weight with respect to the current error.
2. Update the weights using the derivative information, via some form of minimisation scheme to minimise the error figure.

The total error for the training set is then compared to some pre-set tolerance and, if not met, the whole training set is again fed through the net. This is termed one *epoch*, or *iteration*. It typically takes a few hundred epochs for DONNET to converge (reach a stable minimum error configuration).

3.3 DONNET Conceptual Model

Each hidden layer node, along with its associated weight connections to the input layer, represents a cutting hyperplane within feature-space. A particular hyperplane is moved through feature-space by changing the values of the weights going into the associated node. At start-up, DONNET creates enough hidden layer nodes to span all possible pairwise separations of the classes. It is the task of the learning phase to position and connect these hyperplanes so as to separate the classes in as effective a manner as possible. The starting position of these hyperplanes has been shown to dramatically affect the speed of convergence of the network to an optimal solution (see Dunne, 1992; German, 1994). The special instance of Fisher's Linear Discriminant for the two class problem (the first canonical variate) (Mardia et. al., 1974) is used to position each pairwise separating hyperplane, as a good approximation to the optimal discriminating position. In terms of the weight-space, we have fitted the network with weights such that the starting error is in the neighbourhood of an acceptable minima.

4.0 Predicting The Learning Ability of DONNET

4.1 Classification Accuracy

With real-world GIS classification problems, it is unrealistic to expect 100% classification accuracy on all data-sets. Therefore, once a classifier is trained, the user normally requires some indication of its accuracy, prior to being used on a particular data-set. Two types of accuracy are of interest to the GIS user:

- The ability of the net to learn from the data, i.e. its final ability to classify the training set.
- The ability of the net to generalise, i.e. it's performance on previously unseen data.

Accuracy is normally measured by passing a set of example data through the network and reporting on the number of correctly classified samples. Two points immediately arise from this:

- How to present the classification accuracy to the user.
- How to select an appropriate example set.

Selecting the appropriate example set is a delicate issue that is beyond the scope of this paper. Suffice to say that it is important to maintain statistical independence between the training set and example set used when the generalisability of the network is to be ascertained or measured (Sarle 1994). Here we are not concerned with the generalisability, but with the ability of a given network to learn from a particular data-set.

4.2 Presenting Classification Accuracy

The way classification accuracy is reported is affected by several factors:

- The relative number of samples in each class.
- The number of independent ground-truth sites available.

- The requirement of either overall, or class-by-class accuracy.

Generally, the reporting of neural network classifier accuracy has been a little inadequate. To some extent, this is due to most early neural network research work being done with artificial data-sets, in which the number of samples in each class is constant, as much ground truth as required is available and only *overall* accuracy is required. Subsequent researchers, in an attempt to directly compare their classifiers' accuracies with previous work, have tended to select data and report in a similar manner. Hence the most common form of reporting neural network accuracy is to give the total number of correctly classified samples in the example data-set, usually as a percentage (see Civco, 1993; Kamata, 1993 and Skidmore, 1995 for examples). This figure, (PCC, or percent correct) can be very misleading if the class sizes are not identical. As an exercise, compare Tables 1a and 1b, examples of a confusion matrix. Assume that classes 1 and 2 are very close in feature-space and hence more difficult to distinguish, whereas class 3 is easily separable (for example, lupins, wheat and water respectively). In Table 1a, all class sizes are equal, so the final figure of 43% PCC better reflects the difficulty the network has in separating classes 1 and 2 than in Table 1b (67% PCC), where the majority of samples are from class 3. A better approach is to quote the average per-class accuracy (the Average Normalised Response, ANR). This gives a figure of 43% ANR in both cases.

The ANR figure is useful for comparing the accuracy of different classifiers on the same data-set. However, for data

Table 1a : 3 class confusion matrix, equal class sizes

	Class1	Class2	Class3	Totals
True 1	4	2	4	10
True 2	3	5	2	10
True 3	5	1	4	10

Table 1b : 3 class confusion matrix, unequal class sizes

	Class1	Class2	Class3	Totals
True 1	1	3	1	5
True 2	0	1	4	5
True 3	1	1	18	20

analysis, one obvious problem with using simple single figures such as PCC and ANR is that they do not inform the user about the accuracy in any specific area of the data-set *feature-space*. Reporting each specific class accuracy or simply reproducing the confusion matrix conveys more information to the user on the classifier's performance, which can be used to help assess the confidence one has in the classification in different areas of the data-set.

4.4 Qualitative Analysis Of The Training Set Using The Network

We wish to predict whether a useful classification is achievable with a particular DONNET architecture, fitted with a set of starting weights, on a specific training set. Although this will not give us any indication of the generalisability of the network, it will allow us to decide whether or not the data set and the classification scheme we wish to impose are compatible with the DONNET methodology. We will satisfy ourselves, for the moment, with being able to answer the question:

Will training produce a significant improvement in the classification rate over our initial, or starting classification? To begin, we can make use of the initial weights used to identify difficult classification decisions. In the following, we assume that the costs of misclassification are identical for each class.

DONNET initially constructs the network with as many hidden-layer nodes (corresponding to separating hyperplanes in the feature-space) as needed to do a pairwise separation of every possible pairing of the classes. Hence, if there are four classes, six hyperplanes will be used, as there are six unique pairings of classes possible. In many cases however, one particular separating hyperplane may separate two or more other classes. For example, hyperplane A may be constructed to separate classes 1 and 2, but coincidentally may also separate classes 1 and 3, the task for which hyperplane B was constructed. In a sense, hyperplane B is "redundant" and could be used elsewhere if required, so moving the hyperplane out of this local region of feature-space will not increase the classification error between these three classes.

Pruning algorithms have been developed by both the decision tree and neural network communities for reducing the size of networks (see Brieman et. al., 1984, Dunne et. al., 1992). Dunne et. al. (1992) identifies each pairwise separation task occurring at every hidden-layer node by examination of the activated discriminant score given at the output of the hidden-layer nodes as the training set is passed through the network. This is done after training so that redundant nodes can be pruned from the network (usually a pruning tolerance is set to limit the level of pruning). Generally, these pruned networks do not perform as well as the unpruned network, since the trained network has already optimised its hyperplanes and hence most, if not all, are in use. However, the same algorithms can be used to identify redundant hyperplanes *prior* to training, some of which may not be used at all, or using the terminology of pruning methodologies, *exhibit zero confusion*. These redundant hyperplanes can be used by the network elsewhere in the feature-space, when needed. A task matrix can be constructed, similar to that in Table 2. The tasks listed in the Additional Separations column are tasks performed by that hyperplane as successfully as the hyperplane constructed to handle them as primary tasks. Hence, hyperplane A not only performs its own primary task of separating classes 1 and 2, but also separates classes 1 and 3 with the same or better rate of success as hyperplane B. Hyperplane B is therefore redundant in this localised region of feature-space. Note that this does not necessarily mean it should be pruned.

Consider the earlier three class problem, with 30 samples in the training set (see the confusion matrix of Table 1b). It is obvious from the confusion matrix that the classifier has correctly classified 18 out of 20 class 3 samples, but only 1 out of 5 for classes 1 and 2. We can determine more than this from the matrix, however. The rows are the true class allocations and give the errors-of-omission, or conditional probability distributions. The columns give us the errors-of-commission. So from Table 1b, the classifier has not only classified 18 samples of class 3 as class 3, but also 4 samples from class 2 and 1 from class 1 (reading up column 3). In fact, totaling the symmetric off-diagonal positions will produce figures for the total errors of classification (the misclassification rate) for each pairwise separating hyperplane. Table 3 gives this "boundary misclassification rate" (BMR) as a block diagonal matrix for each of the three separating hyperplanes. It can be automatically derived from the confusion matrix of Table 1b. The AVG column gives the average misclassification for all boundaries associated with the respective class; hence the average BMR of 19% for class 2 is calculated from the summation of all class 2 columns and rows. We can see that the class 1:2 hyperplane is the largest contributor to the ANR classification error and in general, the average figure of 25% shows that the class 1 region of feature-space delineated by the hyperplanes is contributing to the most error.

Table 2: Three Class separation. Here, hyperplane B is redundant

Hyperplane	Primary Class Separation	Additional Separations
A	1:2	1:3
B	1:3	-
C	2:3	-

Table 3: Boundary misclassification rate as a percentage of class sizes

	Class1	Class2	Class3	AVG
Class1	-	+30%	**20%**	**25%**
Class2	-	-	**8%**	**19%**
Class3	-	-	-	**14%**

Some entries in the BMR matrix have a positive or negative sign associated with them. The positive sign indicates that all the errors are errors of omission, the negative signs indicate that all are errors of commission. Non-zero figures that comprise of both (dual-error boundaries) have no sign and are highlighted in bold. These differences are important, as they indicate whether or not a simple movement of the associated hyperplane will reduce the local misclassification or not. Figure 1 shows the differences graphically for the two dimensional case, with the separating hyperplane shown as a heavy black line. In Figure 1a, a single-error boundary is shown for two non-overlapping classes (in feature-space). A simple local repositioning of

the separating hyperplane will reduce this error, ideally to zero, by positioning it in the region between the classes. Figure 1b shows the case where there is overlap of the classes and a single-error boundary has been formed. In this case, the error can be quickly minimised, again by a local repositioning of the hyperplane, which will reduce it to a dual-error boundary. However, if the hyperplane is producing a dual-error figure, as in Figure 1c, then simple movement of the hyperplane will not suffice to reduce the error significantly and the boundary complexity will have to be increased. This can only be accomplished by either varying the summation of the regions (accomplished at the output layer level of the network) alone, or by additionally moving across one or more redundant hyperplanes, to aid in building up the complexity of the surface. The task matrix can identify whether or not there are redundant hyperplanes available.

The BMR matrix can therefore be used to identify problem boundaries and determine if a simple movement of the hyperplane is sufficient. Along with the task matrix, we can also determine whether or not there is sufficient redundancy in the number of hyperplanes to reduce the dual-error boundaries' misclassification. Ideally, we would expect the following behaviour during training, in terms of the BMR matrix:

1. A reduction of single-error boundary figures to zero, or else a transformation to a lower-valued dual-error figure.
2. Dual-error figures to be reduced only if there are sufficient redundant hyperplanes available, or the addition/subtraction functions of the output layer of the network can be further optimised, to increase the complexity of the boundary.
3. Zero error boundaries to remain in the same state.
4. We would not expect a transformation of dual-error to single-error boundaries, as this would not generally reduce the local misclassification rate (see Figure 1c).

The ability of the network to reduce its classification error can therefore be determined from these matrices. A localised repositioning of the separating hyperplane can reduce single-error figures in the BMR matrix. If there are any redundant hyperplanes in one area of the feature-space, they can be shifted elsewhere to be used to model more complex boundaries and so reduce some of the dual-error figures.

4.6 An Example

Table 4 shows a typical confusion matrix for a DONNET classifier immediately prior to training. The data-set is from the Kioloa area of New South Wales, Australia, which has been made available as a NASA pathfinder data-set through the Australian National University in Canberra (Lees and Ritman, 1991). There are 9 floristic-level classes to be delineated. Table 5 is the associated BMR matrix. The classes are of unequal sample sizes, with the class totals in column 10. The initial classification rate (before training) is 52.79% ANR (72.15% PCC). Glancing down the main diagonal of

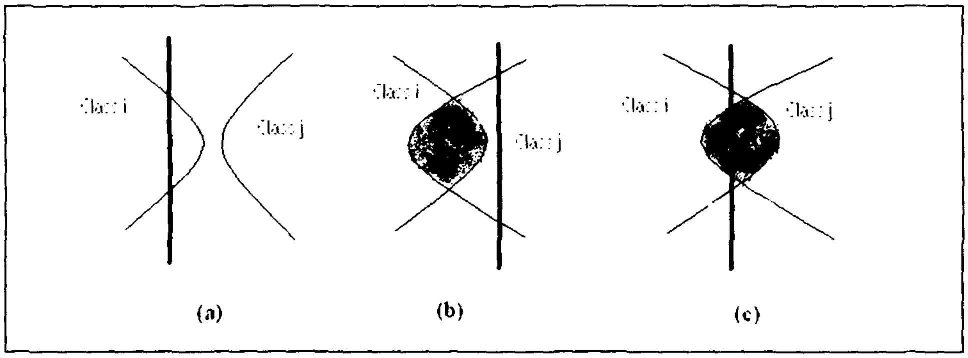


Figure 1: (a) single error boundary, reducible to zero. (b) single error boundary, reducible to dual (c) dual error boundary.

Table 4, we see that the initial network configuration labels classes 1, 4, 5, 7 and 8 quite well. However, there are many errors of omission for class 1, as indicated by its column entries. The conclusion one can draw from this is that the class 1 region delineated in feature-space by the bounding hyperplanes is too large - the hyperplane fit is too loose. In contrast, considering the class 3 column, the hyperplane fit is too tight - all but one example have been excluded from the class.

Analysing the BMR matrix of Table 5, we find that the class boundaries 1:4, 4:5 and 4:6 overlap considerably. Only 11 of the 36 boundaries are dual-error boundaries, there are 12 single-error boundaries and 13 boundaries with no error (overlap) at all. The class 9 feature-space region is correctly delineated; in fact, it is likely that many of the associated hyperplanes are redundant and can be used elsewhere. This applies, to a slightly lesser extent, to class 8 as well and is confirmed by the task matrix, a portion of which is shown in Table 6. There are, in fact, 16 redundant hyperplanes *at the commencement of training* ie. 16 hyperplanes whose primary separation task is performed with the same success by other hyperplanes.

We can conclude, from this rudimentary assessment of the initial task and BMR matrices, that training will give us a significant improvement in classification due to the following:

1. *There are a significant number of single-error boundaries, hence fine-tuning of the associated hyperplane positions should improve these.*
2. There is sufficient redundancy in the number of hyperplanes to aid in the construction of more complex boundaries for some of the 11 cases where we have a dual-error boundary.

Table 4 · 9 Class Confusion Matrix (0 iterations)

	Class1	Class2	Class3	Class4	Class5	Class6	Class7	Class8	Class9	Totals
True 1	**187**	3	0	13	17	3	0	7	0	230
True 2	26	**8**	0	5	2	1	3	5	0	50
True 3	26	2	*1*	*0*	*4*	*1*	3	2	0	39
True 4	39	1	0	**108**	30	0	11	0	0	189
True 5	22	0	0	28	**85**	0	1	0	0	136
True 6	19	0	0	31	2	**18**	3	0	0	73
True 7	17	2	0	14	5	6	**22**	0	0	66
True 8	3	0	0	0	0	0	0	**122**	0	125
True 9	0	0	0	0	0	0	0	0	**374**	374

Table 5 · 9 Class BMR Matrix (0 iterations)

	Class1	Class2	Class3	Class4	Class5	Class6	Class7	Class8	Class9	Avg
Class1	-	**10.4%**	-9.7%	**12.4%**	**10.7%**	**7.3%**	+5.7%	**2.2%**	0%	**7.3%**
Class2	-	-	+2.2%	**2.5%**	-1.1%	-1.6%	**4.3%**	-1.1%	0%	**2.9%**
Class3	-	-	-	0%	-2.3%	-0.9%	-2.9%	-1.2%	0%	**2.4%**
Class4	-	-	-	-	**17.8%**	-11.8%	**9.8%**	0%	0%	**6.8%**
Class5	-	-	-	-	-	+1.0%	**3.0%**	0%	0%	**4.5%**
Class6	-	-	-	-	-	-	**4.7%**	0%	0%	**3.4%**
Class7	-	-	-	-	-	-	-	0%	0%	**3.8%**
Class8	-	-	-	-	-	-	-	-	0%	**0.6%**
Class9	-	-	-	-	-	-	-	-	-	**0.0%**

Table 6 Portions of Task Matrix for 9 Class Example

Hyperplane	Primary Class Separation	Additional Separations
1	1:2	-
2	1:3	1:9, 2:9, 3:7, 3:9, 4:6, 5:7
28	5:7	-
29	5:8	4:8, 6:8
30	5:9	1:6
36	8:9	1:8

In fact, when trained on this data-set for 250 iterations, the network produces the confusion and BMR matrices of Table 7 and Table 8. The classification rate is now 65.66% ANR (with 78.55% PCC). As expected, most single-error boundaries have been reduced to lower valued dual-error boundaries or zero (12 down to 6). Dual error figures have been reduced for all class 1 boundaries, as have all class 4 figures (these were the two classes contributing most to the overall error rate), indicating that boundary complexity has been *increased* around these classes. Again, as expected, no dual-error boundaries have been transformed into single-error boundaries and all zero-error boundaries have remained the same. The overall local misclassification rates associated with class 2 have increased slightly (errors of commission have increased). Class 3 errors are approximately the same and all others have decreased slightly. There are still some single-error boundaries left, so we could expect marginal improvements on the classification rate with further training (but probably at the cost of reduced generalisability).

The procedure for examination prior to training and the information that can be derived is thus as follows:

1. Fit the network with the weights calculated from the discriminant functions.
2. Generate confusion, BMR and task matrices.
3. If there are a significant number of single-error figures

Table 7 9 Class Confusion Matrix (250 iterations)

	Class1	Class2	Class3	Class4	Class5	Class6	Class7	Class8	Class9	Totals
True 1	195	4	2	9	15	3	1	1	0	230
True 2	10	26	0	3	2	1	4	4	0	50
True 3	18	5	7	0	4	0	3	2	0	39
True 4	38	1	0	120	15	4	11	0	0	189
True 5	18	3	0	27	84	1	3	0	0	136
True 6	7	2	1	24	3	34	2	0	0	73
True 7	7	2	0	8	3	3	43	0	0	66
True 8	1	0	0	0	0	0	0	124	0	125
True 9	0	0	0	0	0	0	0	0	374	374

Table 8 9 Class BMR Matrix (250 iterations)

	Class1	Class2	Class3	Class4	Class5	Class6	Class7	Class8	Class9	Avg
Class1	-	5.0%	7.4%	11.2%	9.0%	3.3%	2.7%	0.6%	0.0%	4.9%
Class2	-	-	+5.6%	1.7%	2.7%	2.4%	5.2%	-2.3%	0.0%	3.1%
Class3	-	-	-	0.0%	-2.3%	+0.9%	-2.9%	-1.2%	0.0%	2.5%
Class4	-	-	-	-	12.9%	10.7%	7.5%	0.0%	0.0%	5.5%
Class5	-	-	-	-	-	1.9%	3.0%	0.0%	0.0%	4.0%
Class6	-	-	-	-	-	-	3.6%	0.0%	0.0%	2.8%
Class7	-	-	-	-	-	-	-	0.0%	0.0%	3.1%
Class8	-	-	-	-	-	-	-	-	0.0%	0.5%
Class9	-	-	-	-	-	-	-	-	-	0.0%

in the BMR matrix contributing to the average BMR figures, a worthwhile reduction in error can be expected.

4. [illegible] redundant hyperplanes available, reductions in dual-error figures can be expected.
5. If all error-figures are dual-error and all hyperplanes are in use, no further significant error reduction can be expected.

5.0 Extending the Model

There are two points worth noting about the conceptual model:

- By using the pairwise linear discriminants to calculate the number of hyperplanes and the corresponding initial weights, the net starts from the assumption that the data is linearly separable. This is not as big a restriction as it first seems, if one has at least several output classes. Unless all class centroids are closely clustered within a particular class distribution in feature-space (if this is the case, it is unlikely that any classification strategy will work), at least some of them will be linearly separable. Furthermore, as we have shown, although each hyperplane is linear, complex boundaries can be built up from a combination of hyperplanes, allowing some of the non-linear boundaries to be modelled. The weights between the hidden and output layers allow addition and subtraction of isolated clusters of data, which helps the network classify non-unimodal distributions, as well as aiding in the construction of more complex surfaces from individual hyperplanes.
- As implied above, the complexity of the decision boundaries is directly related to the number of output classes we designate.

The last point is worth further discussion. DONNET may fail to classify a high dimensional feature-space into just two classes, for instance, because of a lack of complexity in the discriminating boundary (in this case, there would be just one hyperplane). Additionally, the situation can arise where all hyperplanes are in use and some of the discriminating boundaries still require a higher level of complexity. These situations can be overcome by the addition of more hidden layer nodes, corresponding to adding more hyperplanes. The problem arises as to how to position these new hyperplanes so as not to disturb the network's position in weight-space too dramatically. The addition of new random weights and nodes to a network will radically shift its position in weight-space, normally with a significant increase in error. If we wish to keep our conceptual model, a simple solution is to re-classify the classes on either side of the hyperplane in question into artificially conceived sub-classes, which increases the complexity available for the decision boundaries. Then we can calculate the discriminant functions for the new weights and continue training, without disturbing the current position of the network in weight-space significantly. We can finally do a sub-class merge to return to the original number of classes.

6.0 Conclusions

The BMR matrix can be used to identify successful discriminating boundaries and those that may require additional hyperplanes. The task matrix can identify any redundant hyperplanes in the system. Production of these matrices can be automated and implemented at any stage of the training phase. Examination of the BMR and task matrices prior to the commencement of training, after fitting the model with the weights derived from the discriminant functions, can show the user where likely improvements can be made to the classification. This examination process can also be automated and can quickly signal the user if there is not enough flexibility in the model to improve the classification rate beyond that produced by the initial conditions. This only takes one pass through the training-set, as compared to the several hundred needed to train the network.

Further work is under-way to automate the spawning of additional nodes when required, as discussed in Section 5.0. Alternatively, the user may wish to stop when the hyperplane redundancy in the network has been depleted. The task matrix can be analysed during training to indicate when this is so. A useful metric for reporting on the spatial distribution of the error, using the error map generated by DONNET should also be developed. This will depend, to a

large extent, on the spatial distribution of the training-set and whether the training sites are isolated points or homogeneous regions across the data-set.

References

Amari S. (1967) Theory of Adaptive Pattern Classifiers, *IEEE Transactions on Elect. Comput.*, Vol. EC-16, pp. 299-307

Bischof H., Schneider W. and Pinz A.J. (1992) Multispectral classification of Landsat images using neural networks, *IEEE transactions on Geoscience and Remote Sensing*, Vol. 30, No. 3, pp. 482-490

Brieman L., Friedman J., Olshen R. and Stone C. (1984) *Classification and Regression Trees*, Wadsworth International

Civco D.L. (1993) Artificial neural networks for landcover classification and mapping, *International Journal of Geographical Information Sytems*, Vol. 7, No. 2, pp. 173-186

Dunne R.A., Campbell R.A. and Kiiveri H.T. (1992) Task Based Pruning, *Proceedings of the 3rd Australian Conference on Neural Networks*, ACNN '92 Melb, pp. 166-169

Dunne R.A., Campbell R.A. and Kiiveri H.T.(1993) Classifying high dimensional spectral data by neural networks, *Proceedings of the 4th Australian Conference on Neural Networks*, ACNN'93 Melb

Gahegan M., German G. and West G. (1996) Automatic Neural Network Configurations for the Classification of Complex Geographic Datasets, *Proc. Int. Conf. on Geocomputation*, University of Leeds, UK, pp. 343-358.

German G.W.H. (1994) Multi Layered Perceptrons and the Classification of Remotely Sensed Data, *Honours Thesis*, School Of Computing, Curtin University of Technology Western Australia

German G. W. H. (1995) The Use of Multi Layered Perceptrons for Remote Sensing Classification with Temporal Data, *Proceedings of the International Conference on Neural Networks*, IEEE ICNN'95 Perth

German G.W.H. and Gahegan M. N. (1996) Neural Network Architectures For The Classification Of Temporal Image Sequences, *Computers And Geosciences*, Vol. 22, No. 9, pp. 969-979

Kamata S. and Kawaguchi E. (1993) A neural net classifier for multi-temporal {LANDSAT} images using spatial and spectral information, *Proceedings of 1993 International Joint Conference on Neural Networks*, IEEE, Vol. 3, pp. 2199-2202

Lees B. G. and Ritman K. (1991) Decision tree and rule induction approach to integration of remotely sensed and GIS data in mapping vegetation in disturbed or hilly environments, *Environmental Management*, Vol. 15, pp. 823-831

Lees B. G. (1994) Decision Trees, Artificial Neural Networks and Genetic Algorithms for Classification of Remotely Sensed and Ancillary Data, *7th Australasian Remote Sensing Conference Proceedings*, pp. 51-60

Mardia K. V., Kent J. T. and Bibby J. M. (1979) *Multivariate Analysis*, London Academic Press

Pao Y. H. (1989) *Adaptive Pattern Recognition and Neural Networks*, Addison-Wesley, Reading, MA

Sarle W. (1994) Neural networks and statistical models, *Proceedings of the Nineteenth Annual SAS Users Group International Conference*, SAS Institute, paper No. 320, pp. 1538-1550

Skidmore A. (1995) Neural Networks and GIS, Australasian Geographic Information Systems Applications, No. 11, pp. 53

Wray B. (1996) A guide to the literature on learning probabilistic networks from data, *IEEE Transactions on Knowledge and Data Engineering*, Vol. 8, No.3 (in press)

The Usher's Approach (UshA) for the Realisation of Fuzzy Clustering Theory for Zonal Analysis in Raster-Based GIS

Thomas Q. Zeng[1], Peter Cowell[1] and Qiming Zhou[2]
[1]Coastal Studies Unit, Department of Geography
University of Sydney, NSW 2006, Australia
[2]School of Geography,
University of New South Wales, NSW 2052, Australia

*Presented at the second annual conference of GeoComputation '97 & SIRC '97,
University of Otago, New Zealand, 26-29 August 1997*

Abstract

This paper investigates the problem of integrating fuzzy clustering theory with raster-based GIS for zonal analysis. A new approach, named the 'Usher's Approach' (UshA), is proposed to provide a solution to the problem. Characterised by its simplicity, yet very practical, this approach has three main advantages: i) it increases the applicability of fuzzy clustering theory in raster-based GIS; ii) it improves the efficiency of data processing and iii) increases overall accuracy of the modeling result.

I. INTRODUCTION

Defined by Lotfi Zadeh (1965), Fuzzy Set theory has proven to be a technique expedient for handling uncertainty, offering the prospect of coping with the inherent complexity in large-scale systems. It provides the basis for automated solutions to an extended range of application problems (Shen and Leitch, 1993). It has been found that fuzzy relation functions can enhance GIS operations in two ways: i) the system can cope with incomplete and even imprecise data in a more user friendly environment; and ii) users are allowed to express their subjective views on the obtained data. As it specially lends itself to representation and reasoning of transitive changes in geographic phenomena, much attention has been paid to the integration of fuzzy set theory with GIS, such as in the FLESS system (Leung and Leung 1993a; 1993b), IDRISI (Eastman, 1993), and the works by Kollias and Voliotis (1991), Sui (1992), Suryana (1993), Brimicombe (1993) Altman (1994), Fisher (1994), soil mapping (Burrough, 1989), land suitability for agriculture and urban settlement (Wang *et al*, 1990), and optimal landfill-site selection (Charnpratheep *et al*, in press).

Although much progress has been made in integrating Fuzzy Set theory with GIS, till now, the fuzzy clustering is only achieved by assigning a membership function to one data layer (or one attribute). Also, no attention has been paid to the use of fuzzy clustering techniques on multi-attributes for zonal analysis in GIS, even though zonal analysis is quite common in environmental modelling. Thus, this paper attempts to fill that niche.

In the context of this study, zonal analysis applies to the specified environmental zone, or area of interest (AOI); for example, the coastal zone, a river buffer area, flora patches, or site selection. The term 'Usher Approach' is coined as an analogy to facilitate explanation of data processing by the methods discussed in this paper.

II. FUZZY CLUSTERING THEORY

Statistically, classification is to cluster the elements of a set into subsets based on their similarity according to certain criteria. Clustering includes two aspects: determining group of elements; the clusters of the groups and determining the relationship of each element to the group. Methodologies have been developed for clustering, such as using Kohonen algorithms (Pal, 1993), Neural networks

(Anand, 1995; Kasabov, 1996) and Fuzzy clustering (Dobois, 1980; Zimmermann 1991; Terano *et al.* 1992; Bazdak 1992, and Kasabov, 1996).

Among these clustering methodologies, Fuzzy clustering aims to achieve a classification that is closer to the real-world, because the object itself is usually of ambiguous, or fuzzy nature. Multi-attribute fuzzy clustering is illustrated conceptually in Figure 1. This paper focuses on how to integrate fuzzy clustering techniques with raster-based Geographical Information Systems (GISs). For fuzzy clustering theory, the reader can refer to works mentioned above.

Fuzzy clustering involves three steps:
i) normalising data into the range [0, 1];
ii) calculating a similarity matrix; and iii) grouping.

i) Data normalisation

Data sets are normalised using Equation (1)

$$x_i' = \frac{X_i - X_{min}}{X_{max} - X_{min}} \quad \text{........(1)}$$

where x_i' is a namorlised attribute value of a pixel, x_i is the attribute value of corespoding pixel in a data layer. x_{min} and x_{max} are *the minimum and maximum value in the grid* data set respectively

The normalisation is applied to all the attribute data layers, then, followed by calculation of similarity matrix.

ii) Calculation of similarity matrix

There are many algorithms for calculating the fuzzy similarity matrix (e.g. Zimmermann, 1991). The vector-multiplying method is used in this study. A fuzzy similarity matrix **ri,j** is created by calculating every other two pixels x_i, x_j in considering a given attribute value *k*, e.g.:

xi 1, *xi* 2, ..., *xi* k

xj 1, *xj* 2, ..., *xj* k

The **ri,j** matrix is defined as:

$$r_{i,j} = \begin{cases} 1 & (i = 1) \\ \dfrac{1}{M}\sum_{k=1} x_{ik} * x_{jk} & (i \neq 1) \end{cases}$$

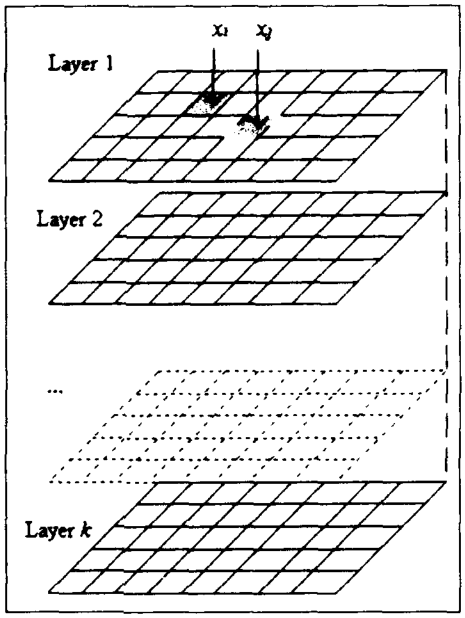


Figure 1 Fuzzy Clustering of Multi-Data Layers

where *M* is a constant.

Equation 3 is illustrated in matrix form by the following:

$$R = r_{ij} = \begin{array}{c|cccccc} & x_1 & x_2 \dots & x_i \dots & x_j \dots & & x_{m\times n} \text{(2)} \\ x_1 & 1 & 0.80 & 0.76 & 0.91 & & 0.88 \\ x_2 & 0.80 & 1 & 0.68 & 0.95 & & 0.79 \\ \vdots & & & & & & \\ x_i & 0.76 & 0.68 & 1 & 0.77 & \dots & 0.58 \\ \vdots & & & & & & \\ x_j & 0.91 & 0.95 & 0.77 & 1 & & 0.83 \\ \vdots & & & & & & \\ x_{m\times n} & 0.88 & 0.79 & 0.58 & 0.83 & & 1 \end{array}$$

Before Fuzzy clustering, the matrix *ri,j* must be converted into a transitive form, that is:

$$R^2 = ri,j \quad \text{..................(3)}$$

where

$$r_{i,j} = \int_{k=1}^{n} (r_{ik} \mid r_{jk}) \qquad [i, j = 1,\ 2, \dots, m;\ k = n]$$

The calculation continues until $R^2 = R^{2n}$ (n = 1, 2, ...)

iii) Grouping

When $R^2 = R^{2n}$, a value of [1, 0] is selected starting from 1 and decreasing to 0. All the pixels, $x_i x_j ... x_{(m \times n)}$ are grouped into classes.

Even though the built-in Fuzzy membership function is available in some commercial raster-based GIS software, such as IDRISI(Eastman 1993), the Fuzzy clustering techniques described above have not been integrated with raster-based GIS. The FUZZY function in IDRISI does not involve multi-layer fuzzy clustering as shown in Figure 1. To-date, most of these fuzzy clustering applications are only associated with a few data-elements (or points), not with the massive grid data sets, due to problems discussed in the next section.

III. THE PROBLEM OF INTEGRATING FUZZY CLUSTERING THEORY WITH RASTER-BASED GIS

As Equations (2) and (3) show, the core of the Fuzzy clustering method is the calculation of the fuzzy similarity matrix. The more numerous the data elements (or points) become, the larger the matrix. The size of the matrix increases exponentially with increasing numbers of the pixel space (Figure 2). For example, if a grid data-set covers an area of 22 x 21 km with a pixel size of 20 x 20 m, giving 1100 x 1050 pixels, the size of the fuzzy similarity matrix would be $(1100 \times 1050)^2 \times 2 = 2.66805 \times 10^{13}$. Since it takes 4 bytes to record the value of each cell, the matrix would entail 1.06722×10^{13} bytes (10,672.2 GB!!!). Considering that the fuzzy similarity matrix is symmetric, i.e. the value for x_i corresponding to x_j is the same as x_j corresponding to x_i, it still requires half size of the memory (i.e. 5,336.1 GB). Experiments undertaken in this study show that the required memory volume increases with power two of the increase in the number of pixels (Figure 2). This means that doubling the spatial resolution of the pixels (e.g. from 20 x 20 m to 10 x 10m) would require an increase in computer memory by eight orders of magnitude.

Therefore, the memory of the hardware is an obstacle to the application of fuzzy clustering theory of multi-attribute analysis in raster-based GIS (Zeng, 1991).

Two conventional approaches exist to overcome this problem. One is to reduce the data volume through degrading the spatial resolution of the data set; i.e., to reduce the numbers of pixels by increase the pixel size, sacrificing accuracy. The other approach is to subdivide the data set into smaller files, e.g. tiling an image, though this results in an increase in CPU time as well as cost. Experiments show that it takes more than 38 CPU hours to process a data set, containing 7 data-layers (covering 22 X 21 km with pixel size of 400 x 400 m,) for fuzzy calculation using a Pascal program on VAX machine (Zeng, 1991). Fuzzy classification of higher resolution data set could not be done

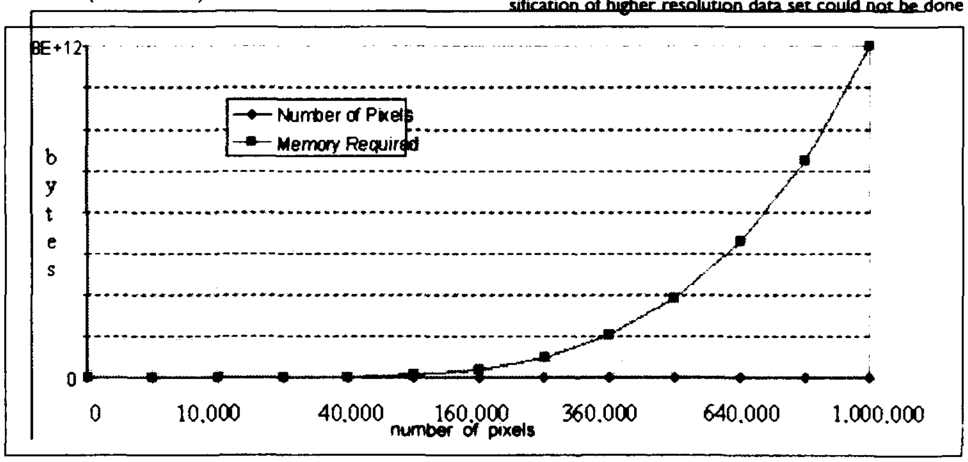


Figure 2. Memory Required vs Number of Pixels

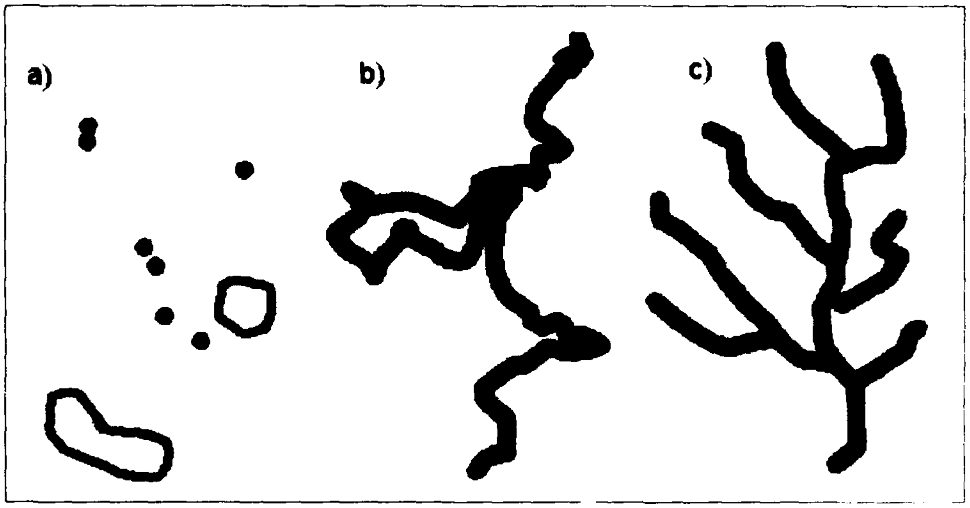


Figure 3. Spatial Patterns of Zonal Analysis. a) Flora Patches, b) Coastal Zone and Lagoon, c) River Buffer Area

due to limitations of memory resources and time constraints in a real-world project. Even though if it could be done, it would be very slow and costly.

A variety of methods for compressed image storage are available, such as Quadtree Decomposition (QD) (Fisher, 1995), and *image partitioning* (Eshaghian, 1991). However, these methods require complex block classification (segmentation) prior to image compression and are of more relevance to image processing. For environmental studies, the interest is usually focused on only parts of the whole data set for a study area. For example, a study of the coastal impact of climate change focuses on phenomena along the coast zone (Cowell & Thom, 1994, Cowell *et al.*, 1995, 1996). Anything outside the coastal zone is not considered (or only as an external factor). This area of interest (AOI) is referred to as the "zonal area" as illustrated in Figure 3. Therefore, an Usher's Approach is proposed and described in next section.

IV. CONCEPTS AND PROCEDURES OF THE USHER'S APPROACH (UshA)

1. Basic Concept of the Usher's Approach.

For zonal analysis, the area outside the AOI is ignored, thereby, reducing the data volume (Figure 3). The *theoretical concept behind the Usher's Approach is in projection* (set transformation) of the basic axioms of traditional set theory (Jech, 1978). Given a universal set $U(x_{ij})$, there is a subset $A(x_n)$ that satisfies the following conditions

$$U(x_{ij}) \Rightarrow A(x_n)$$

$$U(x_{ij}) \supseteq A(x_n) \quad \text{.........(4)}$$

This projects a three dimensional, spatially referenced attribute data set $U(x_{ij})$ into a one dimensional, aspatial subset $A(x_n)$ which meets the requirement that $A(x_n)$, the subset of $U(x_{ij})$, can subsequently be used in the fuzzy clustering calculation. The resultant data set *$A'(x_n)$ is then* converted back to a spatially referenced data set $U'(x_{ij})$ (Figure 4). During the data processing, the individual value of the data set is extracted and stored in the new subset $A(x_n)$. When the fuzzy calculation is completed for each x_{ij}, the resultant data $A'(x_n)$ is then restored into its position one by one, in the same way as an usher in a theatre works. Therefore, the procedure is termed *Usher's Approach* (UshA). The procedure is described below.

2. The Procedure of UshA

The first step is to utilise other GIS function to separate the data set into two types of area in the same raster file, as demonstrated in Figure 4. The areas that lie outside the AOI are assigned a value of zero, with the remaining zones then subjected to be *fuzzy-clustered. Figure* 5 shows the

steps of UshA.

V. DISCUSSION

Figure 6 illustrates coastal risk assessment by the fuzzy clustering using UshA approach. In the UshA method, the extracted subset is:

$A(x_n) = U(x_{ij}) - U(0)$

where U (0) is the total number of pixels that is assigned to zero.

The memory required is:

$$\begin{aligned} M &= \frac{1}{2}k * A(x_n)^2 \\ &= \frac{1}{2}k * [U(x_{i,j}) - U(0)]^2 \\ &= \frac{1}{2}k * [U(x_{i,j})^2 - 2 * U(x_{i,j}) * U(0) + U(0)^2] \quad \text{.....(5)} \end{aligned}$$

where k is a factor depending on the data type, e.g. for integer $k = 2$.

Some GIS software has a 'masking' function that assigns zero area of the region outside the AOI, for effective management of database; e.g., 'SETMASK' command in the GRID module in Arc/Info (ESRI, 1995). However, the zero areas still take some memory. Other techniques have been developed, such as Fractal Image Compression (Jacquin, 1992; Barnsley and Hurd, 1993; Fisher, 1995); and Parallel Pipelined Fractal Compression using Quadtree Recomposition (Jackson and Mahmoud, 1996). These techniques basically deal with a single image to define a quarto-structure for homogeneity pixels of an image, in which there are some trade-offs in image quality. In environmental studies, particularly in multi-data layer analysis, the information is usually kept as high-resolution as possible, for each individual pixel in the multi-data layer. If the conventional methods discussed in Section III are used, then the resolution of a more detailed data layer must be reduced to match with other lower resolution data layers, due to the memory problems. This diminishes overall accuracy of the analysis. In contrast, the UshA method maintains the high quality of the data-layers. Therefore, UshA is considered a more appropriate approach if the full potential fuzzy clustering theory in GIS is to be realised.

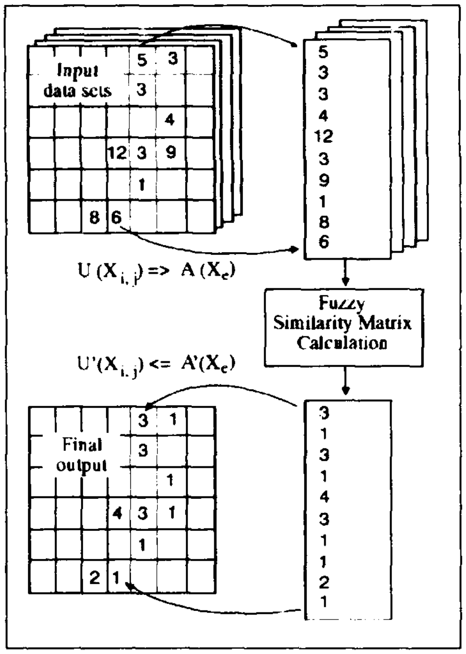


Figure 4. Usher's Approach

VI. CONCLUSIONS

The Usher Approach (UshA) has the following advantages:

i) UshA makes it possible to implement fuzzy clustering theory within GIS for zonal analysis, which was previously impracticable.

ii) UshA significantly reduces memory requirements and improves the efficiency of data processing, saving time and cost.

iii) UshA increases accuracy in both spatial and aspatial terms. Because data outside the AOI can be regarded as 'noise', the UshA method allows only those data of interest to be processed. In this way, the error due to the 'noisy' data is minimised and accuracy will be increased. On the other hand, if some layers of a database have higher spatial resolutions than the others, then the lower spatial resolution can be interpolated to increase the spatial resolution while the higher resolution remains; the higher the resolution of a data set, the higher the accuracy of the analysis result.

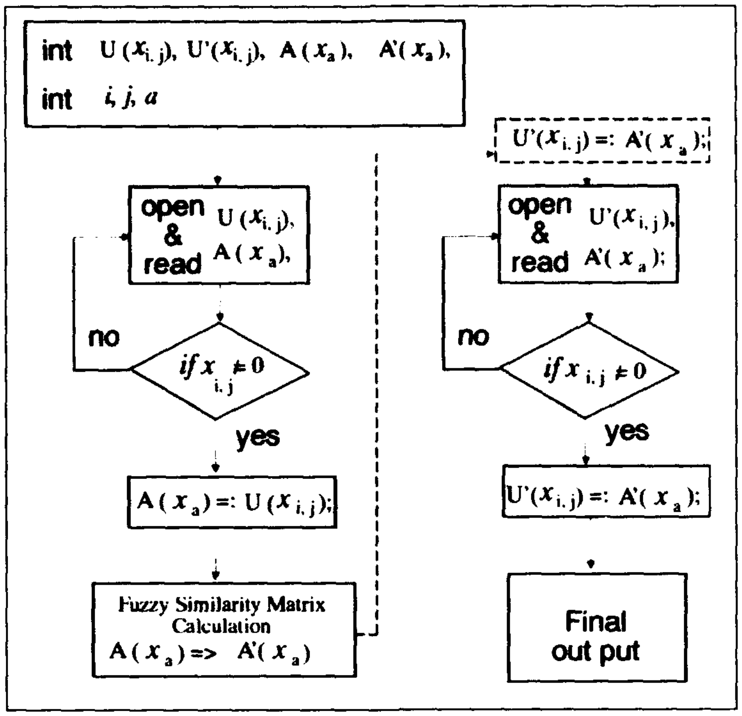


Figure 5. Procedures of Usher Approach

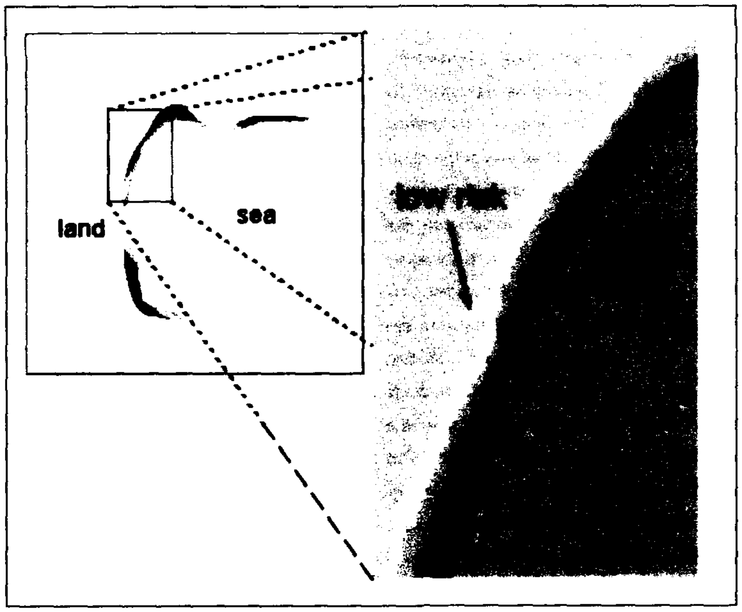


Figure 6. Illustration of UshA for Coastal Hazard Assessment

The UshA provides a methodology for realising the integration of fuzzy cluster theory with raster-based GIS. From the view-point of environmental studies, in which GIS plays an ever-increasing role, the spatial patterns of zonal analysis shown in Figure 3 are quite common. The application of the UshA is prospective and the technique can be applied to many problems in environmental science, such as bio-environmental studies, non-point source pollution, wildlife corridors, as well as assessment of coastal vulnerability due to climate change.

Acknowledgments

The authors would like to thank to Mr Justin Meleo, of Coastal Studies Unit, University of Sydney, and the reviewers for their critical review of this paper. This research is supported by the Australian National Greenhouse Advisory Committee and is contribution to the European Commission MAS III project, PACE.

REFERENCE

Altman, D. (1994). Fuzzy Set Theoretics Approaches for Handling Imprecision in Spatial Analysis. Int. J. Geographical Information Systems, **8(3)**, 271-289.

Anand, R. (1995). Efficient classification for multiclass problems using modular neural networks. *IEEE Transactions on Neural Networks* v. 6, p117-24.

Barnsley, M.F. and Hurd, L.P.(1993): *Fractal Image Compression*. AK Peter Ltd, Wellesley, MA.

Bezdek, J. C. 1992: Numerical convergence and interpretation of the fuzzy c-shells clustering algorithms. *IEEE Transactions on Neural Networks* v. 3 (Sept. '92) p787-93.

Brimicombe, A. J. (1993). Combining Positional and Attribute Uncertainty Using Fuzzy Expectation in a GIS. In GIS/LIS'93, Minneapolis, Minnesota: Vol. I (pp. 72-81).

Burrough, P. A. (1989). Fuzzy Mathematical Methods for Soil Survey and Land Evaluation. J. of Soil Science, **40**: 477.

Charnpratheep, K.; Zhou, Q. and Garner, B. Preliminary landfill site screening using fuzzy Geographical Information Systems. *Waste Management and Research*, (in Press, 1996)

Cowell, P.J. and Thom, B. G. (1994a) Coastal impacts of climate change - modeling procedures for use in local government. *Proc. 1st National Coastal Management Conference, Coast to Coast ' 94*. Hobart, June. 43-50.

Cowell, P.J, Roy, P.S., Zeng, T.Q. and Thom, B.G.(1995): Relationships for predicting coastal geomorphic impacts of climate change. *Proc. Int. Conf. Ocean and Atmosphere Pacific*, Adelaide, October 1995, 16-21.

Cowell, P.J, Roy, P.S., Zeng, T.Q., Hennecke, W., and Thom, B.G.(1995): Regional predictions of climate change impacts. *Proc. Australian Coastal Management Conference, Coast to Coast ' 96*.

Dobois, D., (1980): *Fuzzy sets and systems theory*, Academic Press, New York.

Eastman, J.R. (1995): *IDRISI for Windows*, User's Guide, Version 1.0, (Worcester, Massachusetts: Clark University. p8-20.

Eshaghian, M.M.(1991): Parallel algorithms for Image Processing on OMC. *IEEE Trans. on Computers*, **40**(7):827-33.

ESRI(1995): *Cell-based Modeling with Grid, Arc/Info* manual, Environmental Systems Research Institute, Inc., Redlands, USA, 481p.

Fisher, P. F. (1994). Probable and Fuzzy Models of the Viewshed Operation. In M. F. Worboys (Eds.), Innovations in GIS: Selected Papers from the First National Conference on GIS Research UK, London: Taylor & Francis. p161-175.

Fisher, Y. (1995): *Fractal image compression: Theory and Application*, Springer-Verlag, Berlin. 341p.

Jacksno, D.J. and Mahound, W. (1996): Parallel Pipelined Fractal Compression using Quadtree Recomposition. *The Computer Journal*, **39**(2):1-13.

Jacquin, A.E. (1992): Image coding based on a fractal theory of iterated contractive image transformations. *IEEE Trans. Image Process.*, **1**:18-30

Jech, T.J. (1978): *Set Theroy*, Academic Press, New York, 1978, 621p.

Kasabov, N. K. (1996): *Foundations of neural networks, fuzzy systems, and knowledge engineering.* Cambridge, Mass. : MIT Press, 1996, 550 p.

Kollias, V. J., & Voliotis, A. (1991). Fuzzy Reasoning in the Development of GIS. Int. J. Geographical Information Systems, **5**:209-223.

Leung, Y. and Leung, K.S. (1993a) An intelligent expert system shell for knowledge-based Geographical Information systems: 1 the tools. *International Journal of Geographical Information systems*, **7**:189-199.

Leung, Y. and Leung, K.S. (1993b) An intelligent expert system shell for knowledge-based Geographical Information systems: 2 some applications. *International Journal of Geographical Information systems*, **7**:201-213.

Pal, N. R. (1993). Generalised clustering networks and Kohonen's self-organizing scheme. *IEEE Transactions on Neural Networks* v. 4,p549-57.

Shen, Q. and Leitch, R (1993) Fuzzy Qualitative Simulation. *IEEE Transactions on Systems, Man, and Cybernetics*, **23**(4):1038-1061.

Sui, (1992) A fuzzy GIS modelling approach for urban land evaluation. Computers, Environment and Urban Systems, **16**:101-115.

Suryana, N. (1 93) The application of fuzzy sub-set theory using certainty factor(CF) in handling imprecision data handled by Geographical Information Systems (GIS), in J. Harts, H.F.L. Ottens and H.J. Scholten (Eds.), EGIS'93, Genoa, Italy;Vol.2, 1325-1337.

Terano, T., Kiyoji, A., Sugeno, M. (1992). *Fuzzy systems theory and its applications* Boston :Academic Press, 1992, 268 p.

Wang, F., Hall, G.B. and Subaryono (1990) Fuzzy information representation and processing in conventional GIS software: database design and application. *International Journal of Geographical Information Systems*, **4**: 261-283.

Zadeh, L.A. (1965) Fuzzy Sets. Information and Control **8**:338-353.

Zeng, T. Q. (1991) A fuzzy set analysis approach to environmental geomorphological mapping: Integration of remotely sensed data and other data with geographical information systems. *Tropical Geomorphology.* (South China Normal University, Guangzhou, China) **15**(2):24-72.

Zimmermann, H.-J. (1991). Fuzzy Set Theory-and Its Applications (2 ed.). Massachusetts: Kluwer Academic Publishers. 399p.

Contents

Foreword

The second GeoComputation Conference (GeoComp 97) and the 9th Annual Spatial Information Research Centre Colloquium (SIRC 97) have coalesced at Otago in 1997. It is an appropriate advance that the University of Leeds and the University of Otago combined these two events which are having an increasing impact on the geocomputing and spatial analysis communities.

GeoComp 96 was held in Leeds and was a great success and 97 continues the tradition. Welcome to the vibrant provincial city of Dunedin and a very warming welcome to the University of Otago. We are pleased you are here and trust that the conference will be rewarding.

The conference consists of over 40 research papers that are either presented orally or as a poster. All papers are printed in these proceedings and are available in a variety of electronic forms - namely CD and eventually on the conference web site. We have taken seriously our obligation to 'spread the word'. We have made this concerted effort for a number of reasons. First, it is important that subsequent GeoComp conferences are successful and this conference acts as an advertising agent. Second, we believe it is important to test and report research developments to the industry and for our peers to evaluate our work. Third, Otago, like Leeds, makes a valuable contribution in spatial and geocomputational research and we want you to know about it. Otago and Leeds are proud to bring GeoComputation 97 and these proceedings to you.

There is a very broad spectrum of papers and subjects presented in these proceedings, and furthermore the authors hail from many far flung corners of the globe. The research is truly international. The themes that bind the conference are environmental modelling, artificial intelligence techniques, spatial modelling, integration of geographical analysis tools, cellular automata and visualisation. All these together form a compelling research area - geocomputing. The two additional outstanding themes - important for their predicted omnipresence are distributed environments and data analysis. These two alone will push the capabilities of geocomputing to the existing limits - and beyond.

The proceedings are brim full of the latest ideas and techniques. The authors have toiled hard to present their ideas and the editors have spent many long and sleepless nights getting the bound copy to you . Read and enjoy it.

See you at Otago now, and sometime, somewhere in the future.

George L Benwell
August 1997

Author Index

Some Initial Experiments with Neural Network Models of Flood Forecasting on the River Ouse

L. See, S. Corne, M. Dougherty' and S. Openshaw
School of Geography, University of Leeds, United Kingdom
'Centre for Research on Transportation and Society, Hogskolan Dalarna, Sweden
E-mail: L.See@geog.leeds.ac.uk

Presented at the second annual conference of GeoComputation '97 & SIRC '97, University of Otago, New Zealand, 26-29 August 1997

Summary

This paper presents results and conclusions from a set of experiments designed to assess the potential for using artificial neural networks in real-time flood forecasting, and which highlight the principal benefits and limitations of using the technology in this context. The emphasis on hybrid approaches reflects the need to integrate existing conventional methods with alternative artificial intelligence techniques in order to produce better and more cost-effective forecasting systems.

1 Introduction

Neural networks are widely regarded as a potentially effective approach for handling large amounts of dynamic, non-linear and noisy data, especially in situations where the underlying physical relationships are not fully understood. They are now increasingly being employed to model complex problems as both substitutes for, and in association with, more conventional mathematical and statistical models. Neural networks are also particularly well suited to modelling systems on a real-time basis, and this could greatly benefit operational flood forecasting systems which aim to predict the flood hydrograph for purposes of flood warning and control. Existing flood forecasting models are highly data specific and based on complex and expensive-to-maintain mathematical models. Performance is related to accurate real-time data inputs, the quality of the knowledge used to specify, build and operate the models, and the ability of the models to respond to dynamic and sometimes rapidly changing events. Soft computing approaches, on the other hand, offer real prospects for a cheaper, yet more flexible, less assumption-dependent and adaptive methodology well suited to modelling flood processes, which by their nature are inherently complex, non-linear and sometimes life critical. At a time when global climatic change would seem to be increasing the risk of historically unprecedented changes in river regimes, it would appear to be appropriate that alternative representations for flood forecasting should be considered. However, these new types of models will need to be less dependent on historical data and rely more on real-time adaptation to actual flood events, some of which may be unlike anything seen before. It has also emerged that a critical factor in achieving good model performance is disaggregation of the data. This allows the model builder to draw upon aspects of domain knowledge whilst retaining the advantages of a data driven approach. Hybrid approaches, which supplement conventional methods with artificial intelligence, may well provide both a better and more robust forecasting system, capable of adapting to changing conditions once their construction is better understood and their performance demonstrated on off-line data.

The purpose of this paper is to present a set of empirical experiments which are part of a larger, on-going feasibility study to assess the potential use of neural networks for real-time flood forecasting. A subset of historical water level data from the Ouse River catchment in the United Kingdom was used to build neural network models for two prediction points in the catchment. Assessment is

based on the performance of these models relative to benchmark statistical models and naive predictions evaluated according to goodness of fit, although in practice the performance of flood forecasting systems is based on the number of correct warnings issued and not merely on the accuracy of predicted water levels. Additional customised evaluation measures and the certainty of forecast levels are discussed.

2 Conventional Methods of Flood Forecasting

There are currently two main approaches employed in hydrological forecasting. The first is a mathematical modelling approach. It is based on modelling the physical dynamics between the principal interacting components of the hydrological system. In general, a rainfall-runoff model is used to transform point values of rainfall, evaporation and flow data into hydrograph predictions by considering the spatial variation of storage capacity. A channel flow routing model is then used to calculate water movement down river channels using kinematic wave theory. A snowmelt model is also customary in colder climates. An example of this type of deterministic modelling is the River Flow Forecasting Model (RFFS), a large scale operational system currently employed on the Ouse River catchment (Moore *et al.*, 1994).

The second main approach to flood forecasting is modelling the statistical relationship between the hydrologic inputs and outputs without explicitly considering the physical process relationships that exist between them. Examples of stochastic models used in hydrology are the autoregressive moving average models (ARMA) of Box & Jenkins (1976) and the Markov method (Yakowitz, 1985; Yapo *et al.*, 1993). ARMA models work on the assumption that an observation at a given time is predictable from its immediate past, i.e., a weighted sum of a series of previous observations. Markov methods also rely on past observations but the forecasts consist of the probabilities that the predicted flow will be within specified flow intervals, where the probabilities are conditioned on the present state of the river.

3 Artificial Neural Networks

Artificial neural networks offer a significant departure from the conventional approach to problem-solving and have been applied successfully to a variety of application areas including pattern recognition, classification, optimisation problems and dynamic modelling. Although neural networks were historically inspired by the biological functioning of the human brain, in practice the connection is more loosely based on a broad set of characteristics which they both share, such as the ability to learn and generalise, distributed processing and robustness; see Openshaw & Openshaw (1997) for an overview of neural networks.

The basic function of a neural network, which consists of a number of simple processing nodes or neurons, is to map information from an input vector space onto an output vector space. These neurons are distributed in layers which can be interconnected in a variety of architectural configurations. Information is delivered to each neuron via the weighted connections between them. Information processing within each neuron normally comprises two stages. In the first stage, all incoming information is converted into an activation, where the most common activation function is simply the sum of the weighted inputs. In the second stage, a transfer function, such as the sigmoid, converts the activation into an output value.

A neural network learns to solve a problem by modifying the values of the weighted connections through either supervised or unsupervised training. In supervised training, neural networks are provided with a training set consisting of a number of input patterns together with the expected output, and adjustments to the weights are based on the differences between observed and expected output values. These adjustments are calculated using a gradient descent algorithm (optionally modified for higher performance with refinements such as a momentum or second derivative term). Currently, the most widely applied network is the multilayer perceptron using a supervised training algorithm, known as backpropagation. Once trained, the network is validated with a testing dataset to assess how well it can generalise to unseen data. In unsupervised training, the artificial neural network attempts to

identify relationships inherent in the data without knowledge of the outputs and is often used in classification problems.

To date there have only been a handful of neural network applications that address the hydrological forecasting problem. Research has shown that neural networks have great potential as substitutes for rainfall-runoff models (Abrahart & Kneale, 1997; Minns & Hall, 1996; Smith & Eli, 1995), and Yang (1997) has demonstrated the success of neural networks, trained with a genetic algorithm, in predicting daily river levels on the Yangtze River at Yichang in China. The question now is how well can they perform when asked to make short-term predictions of river flow.

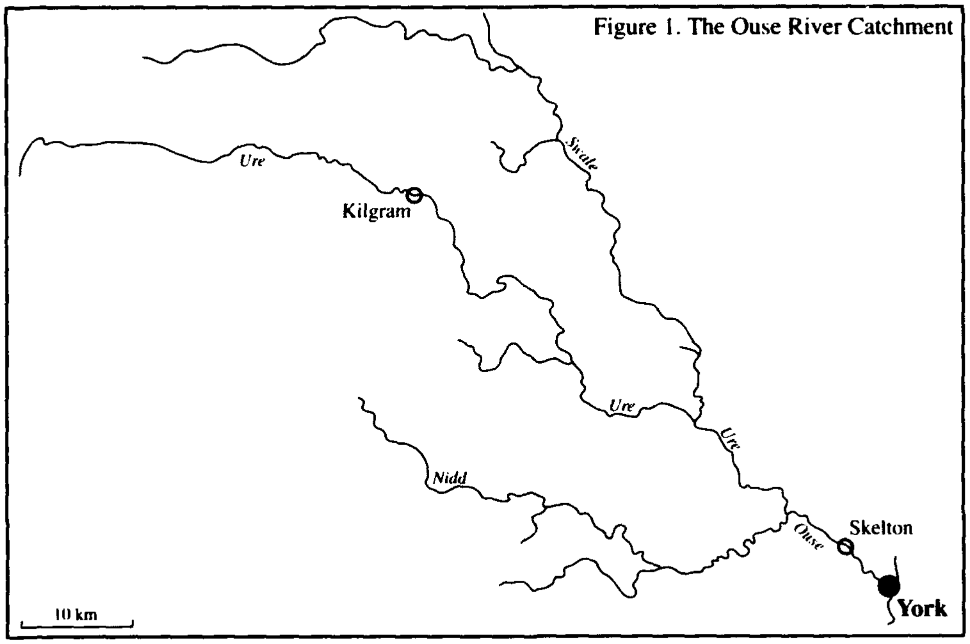
4 Empirical Experiments

4.1 Prediction Points and Forecasting Horizons

The Ouse River catchment in Northern England is subject to seasonal flood events and is the focus of attention because of the availability of historical data. It is fairly typical of a UK catchment, with a mix of urban and rural land uses, and is 3286 km^2 in size. There are three main tributaries: the Nidd, the Swale and the Ure. Gauging stations are distributed throughout the catchment, on each of the tributaries that flow into the River Ouse toward the city of York. Two gauging stations were chosen as prediction points: Skelton, located just north of York on the River Ouse, and Kilgram, located further upstream from York on the River Ure. A map of the study area is given in Figure 1. Due to its location far from the headwaters, Skelton has a relatively stable regime while Kilgram, situated further upstream and hence closer to the headwaters, has a regime that is flashier, with corresponding flood types that are more difficult to predict. All data were originally recorded at 15 minute intervals but to reduce the amount of data and thereby determine whether coarser resolution data were sufficient for prediction, hourly averages were used. No data were missing in this subset.

For prediction purposes, operational forecasting horizons of six and twelve hours are needed. The length of time required for the practicalities of flood protection (e.g., alert-

Figure 1. The Ouse River Catchment



ing the police, issuing of warnings to industries and households in the vicinity, protection of property, etc.) precludes lead times of less than six hours. In fact, a six hour forecast is generally too short for large scale floods but in real-time operational flood forecasting, the catchment is monitored continually and accurate six hour predictions would still be a useful source of management information. More critical is an accurate twelve hour forecast which would allow flood control teams to respond to imminent flood conditions and operate warning systems for the public as well as industries. Water levels of 3 and 2.5 metres at Skelton and Kilgram, respectively, currently trigger standby alarms to duty officers monitoring the catchment. If the levels continue to increase, site specific operational instructions are issued such as shutting flood gates and other engineering measures. In parallel, warnings of varying severity related to flood risk are issued to authorities who inform residents and businesses in the affected areas.

4.2 Establishing Benchmarks

Neural network models need to be tested and compared with benchmarks provided by conventional methods. An initial benchmarking exercise was undertaken to assess the performance of the RFFS model (MAFF Project OCS967P, 1997), currently used by the Environment Agency. The RFFS was used to predict five flood events at forecast horizons of six, twelve, eighteen and twenty-four hours for two stations: Viking, located at York, and Boroughbridge, situated upstream. The results showed that forecast performance degraded appreciably with an increase in forecast horizon, indicating that neural networks could be used to supplement forecasts at these longer time scales. Since the RFFS has not yet been configured to output historical forecast data, direct statistical comparison was not possible. In the future, as performance is evaluated on additional measures available as model output, including peak prediction and time-to-peak, additional benchmark runs will be undertaken at all prediction points for a comprehensive comparison.

For this exercise, benchmarks were produced for a six and twelve hour forecast at each station by (1) fitting ARMA models to the data, and (2) by making naive predictions. Naive predictions substitute the last known figure as the current prediction and are a good bottom line benchmark. ARMA models use a weighted linear combination of previous values and shocks. Five years of hourly water level data (1982-1986) using the measurement at time t and the previous five hourly observations as inputs were used to fit ARMA models to the data using software developed by

Table 1: RMS errors in metres for the ARMA models and naive predictions

		ARMA		Naive Prediction	
Station	Levels	Testing	Validation	Testing	Validation
Skelton	All levels	0.119	0.094	0.187	0.167
6 hr	Low flows	0.082	0.067	0.174	0.154
prediction	Alarms	0.228	0.138	0.476	0.437
Skelton	All levels	0.240	0.201	0.341	0.309
12 hr	Low flows	0.181	0.166	0.319	0.286
prediction	Alarms	0.622	0.498	0.926	0.862
Kilgram	All levels	0.194	0.184	0.299	0.275
6 hr	Low flows	0.131	0.137	0.194	0.195
prediction	Alarms	0.795	0.734	0.954	0.864
Kilgram	All levels	0.224	0.204	0.329	0.300
12 hr	Low flows	0.203	0.187	0.293	0.271
prediction	Alarms	1.188	1.094	1.329	1.202

Masters (1995). Three years of testing data (1987-1988) were then used to validate the model performance. Table 1 lists the RMS errors of the ARMA models and naive predictions for each station and forecasting horizon. In addition, RMS errors are given for low flow levels and levels above which the initial alarms are triggered according to the operational definitions developed by the Environment Agency (1996) in forecasting floods on the Ouse catchment (1996).

The following observations can be made from the model results given in Table 1:

a) the errors for Kilgram are higher then Skelton, reflecting the flashier nature of the upstream station;
b) levels for the longer forecasting horizon are more difficult to predict at both stations, although the difference is less pronounced for Kilgram. This may reflect the generally lower observed levels at Kilgram than Skelton;
c) low flows are much easier to predict than high ones yet it is the flood events which trigger alarms and warnings that are of greatest importance;
d) naive predictions are 30 to 40% worse on average than the ARMA model forecasts for all levels; and,
e) validation results are generally lower than the training results indicating that there must be fewer storm events or storms of a lower magnitude in the validation dataset. Therefore, alternative performance measures are needed in addition to global goodness of fit statistics which can be used to assess overall flood-related performance rather than performance averaged out on all levels.

Neural networks could be used to improve the forecasts of the higher flow levels and the longer forecasting horizons where the performance of the ARMA models is the poorest.

4.3 Neural Network Models

A feedforward backpropagation neural network model was initially developed for predicting levels at Skelton with a six hour forecasting horizon using the Stuttgart Neural Network Simulator package (SNNS Group, 1990-95). The same data inputs as used for the ARMA model and naive predictions were employed for testing and validation of the neural network. A variety of different architectures was examined, and the best result was an overall RMS error of 0.108 metres, obtained with a fully connected multilayer neural network containing 24 neurons in each of two hidden layers. The network was trained for about 17,000 epochs using a momentum of 0.5 and a gain of 0.2. An improvement of just over 9% relative to the ARMA model was obtained. However, after disaggregating the data into low levels and flood events, the results indicated that all the neural network improvements over the benchmarks were gained on the low level events, and the ARMA model and naive predictions outperformed the neural network at the higher levels. Given that the low level events comprised more than 90% of all observations in the dataset,

Table 2: RMS errors in metres for high level events for both the neural network and ARMA models

Station	Levels	Testing	Validation
Skelton	ARMA	0.228	0.138
6 hr prediction	Neural net	0.129	0.118
Skelton	ARMA	0.662	0.498
12 hr prediction	Neural net	0.343	0.280
Kilgram	ARMA	0.795	0.734
6 hr prediction	Neural net	0.321	0.370
Kilgram	ARMA	1.187	1.094
12 hr prediction	Neural net	0.440	0.567

the neural network appeared to be concentrating most learning efforts on low level events and was not able to learn the high level events satisfactorily, regardless of network size or training time. The final trained neural network also seemed to converge on a similar solution to the ARMA model, i.e., both models exhibited similar deviations from the actual observations. In general the deviations were small oscillations about a very smooth observed level record and these oscillations were generally more pronounced at higher levels, accounting for the poorer performance. This indicates that both the ARMA and neural network models are extremely sensitive to small changes in previous measurements. Since the number of low level events is much larger than flood events, both models become good at recognising small changes, but when large changes in level were encountered under a flood situation, the resulting predictions were highly exaggerated. This also indicates that a global model for predicting all river levels is inappropriate.

Therefore, a subset of the data comprising only high level events (as defined above by the levels at which alarms and warnings are triggered), was used to train a feedforward backpropagation neural network for each station and forecasting horizon. Network architecture was reexamined, and smaller neural networks with two hidden layers of 6 and 12 neurons were finally chosen. Convergence varied between 40,000 and 70,000 epochs using a momentum of 0.6 and a learning coefficient of 0.8. Results from the neural networks and the ARMA models are listed above in Table 2.

The results show that there were significant improvements in overall performance of high level event prediction with the neural network, especially for the flashier upstream station and for the longer forecasting horizons. However, given that the input data included only previous measurements, the neural network models were not able to learn those situations where the hydrograph was starting to fall, resulting in peak hydrograph overprediction. This can be seen in Figure 2, which shows flood events over a period of one month in 1988 (part of the validation dataset). The neural network is only predicting higher levels at which alerts and warnings are triggered as defined above, and ARMA predictions are substituted at the remaining lower levels for graphing purposes. Looking at only higher level events, the amount of overprediction is slightly less for the neural network models than it is for the ARMA models. The amount of overpredicton is also more pronounced at longer forecasting horizons. A degree of oscillating behaviour is still apparent by the neural network models at the higher flows, and this behaviour can clearly be seen by looking at the ARMA model predictions at lower levels.

Having established that neural networks can improve performance, additional data inputs such as rainfall and upstream level data appropriately lagged to account for travel times between stations could be incorporated into the simple neural network model. These extra variables should allow the network to learn the rising and falling limbs of the hydrograph more accurately and then more comprehensive evaluation measures such as the peak prediction and the time-to-peak prediction can be employed to assess the performance of the models. As mentioned previously, additional benchmarks from the Environment Agency's own RFFS model will then be used to compare performance.

5 Towards an Operational Hybridised Flood Forecasting System

The initial results of the neural network models for Skelton and Kilgram indicate that there are improvements in performance to be gained by using neural networks as an additional tool for flood forecasting. As a global model, neural networks perform worse than statistical models on the important flood events but by simply disaggregating the data into low and high level events, the neural networks can concentrate their efforts on learning a smaller number of similar patterns and thus significantly improve their forecast accuracy. Experiments with a more intelligent disaggregation scheme have already been undertaken as part of the ongoing feasibility study, whereby previous level measurements, rainfall data and information about

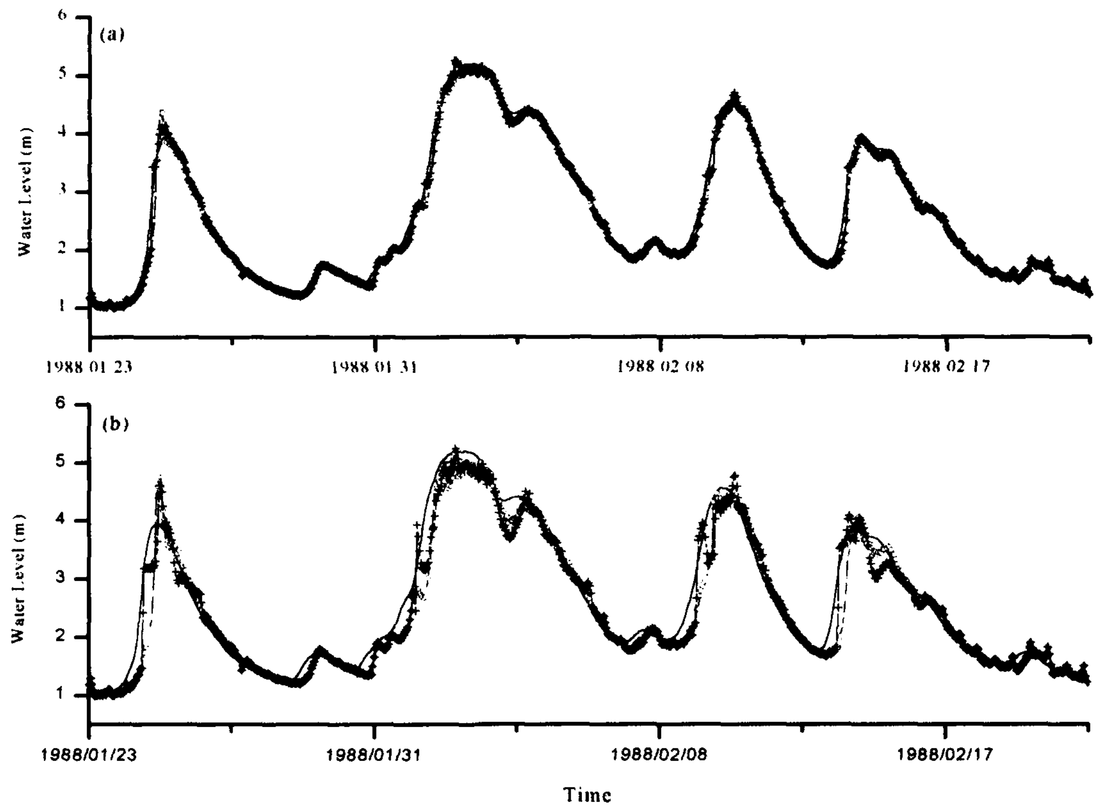


Figure 2 Neural network and ARMA predictions for (a) 6hr and (b)12 hr forecasts at Skelton The observed values are denoted by straight lines. ARMA models by crosses and neural network predictions by open circles

day length have been clustered with a self-organising neural network (Kohonen, 1984). The idea is to produce a series of characteristic event types, for example, low level events with little rainfall, rising hydrograph events with high amounts of rainfall, etc., and build a neural network for each event type similar to the approach undertaken by Van der Voort *et al.* (1996) in forecasting traffic flow. The main advantage of this approach is that each network can concentrate on learning only a small task so that training is quick; moreover, the antecedent conditions of a wet or dry catchment or high levels of evaporation can be incorporated into the models, thereby capturing some of the physical properties of flood events which is analogous to the model states of a large-scale hydrodynamic modelling system. A disadvantage of this approach is the large numb of models which need to be built because the total fore casting will eventually cover 15 prediction points scattered throughout the catchment. However, since training times are quick, which could be a critical factor in adaptive neural networks, overall development time should still be relatively minor compared to the development times associated with large-scale physical hydrodynamic flood forecasting systems. The individual flood prediction models will eventually be linked via a fuzzy logic model that will recommend which sub-network model to use at the current time t based on similar inputs used in the clustering exercise. This is a variation of an approach used by Dougherty (1997) in which a simple Bayesian technique is used to switch between different forecasting methods although the fuzzy logic approach will be a more generalised version of the Bayesian one. When transitions between event types occur, the fuzzy logic model will be able to recommend more than one model but to differing degrees and the resulting prediction will be a weighted average of the suggested models. In this way, the imprecision associated with event types that occur over a broad spectrum of behaviour will be captured directly in the modelling system. Other possibilities include integrating ARMA models and predictions from the conventional hydrodynamic system (RFFS) currently employed by the Environment Agency into the controlling fuzzy logic model to improve forecasts in situations where conventional models may still outperform the neural networks or for low flow forecasting which is less important for flood control.

In addition to improving forecasts, a hybridised system will need to produce certainty measures, i.e., an indication of the certainty of the forecast which operational staff can use in determining whether to issue a flood warning. As PCs and workstations become more powerful, it will be feasible to bootstrap the forecasts and the resulting confidence intervals can then be used as a means of assessing the quality of the forecasts being made. There is no suggestion that human flood managers should as yet be replaced by machines but there is every indication that they may be better able to handle difficult events when aided by automated, adaptive, smart flood forecasting methods can learn to trust.

6 Conclusions

Initial results of an assessment of neural networks for real-time flood forecasting indicate that significant improvements in performance can be gained over conventional statistical models and naive forecasts. A critical factor to good model performance is in disaggregation of the data, as a global model appears to be inappropriate for forecasting all event types. Neural networks in combination with conventional methods and other artificial intelligence technologies have the potential to deliver hybridised solutions that can produce more cost effective and accurate forecasting systems, which can be used as part of a flood warning decision support system. This allows the model builder to draw upon aspects of domain knowledge whilst retaining the advantages of a data driven approach. These neural network hybridised systems are also generic in that they can be widely applied, provided sufficient historic data series are available. Finally, as computer hardware speeds continue to improve, it will be possible to investigate real-time training as flood events proceed, and this may well be expected to yield significant extra benefits.

Acknowledgments

Level data for the Ouse river catchment were kindly provided by the Environment Agency. The research has been

funded by the Ministry of Agriculture, Fisheries and Food under their Open Contracting System under MAFF grant OCS967P. The opinions expressed in this paper are those of the authors and do not reflect Ministry policy.

References

Abrahart, R.J. and Kneale, P.E. (1997). Exploring neural network rainfall-runoff modelling, paper submitted to the *Sixth National Hydrological Symposium*, University of Salford, Manchester, 15-18 September, 1997.

Box, G.E.P. and Jenkins, G.M. (1976). *Time Series Analysis: Forecasting and Control*, Holden-Day, Oakland, California.

Dougherty, M.S. (1997). Meta-analysis of various short term load forecasts. Presented at the *Engineering Applications of Neural Networks* Conference, Stockholm, June 1997.

Environment Agency, (1996). *North East Region Flood Warning Manual Part 1*, Environment Agency, Leeds.

Kohonen T., (1984), *Self-Organisation and Associative Memory*, Springer-Verlag, Berlin.

MAFF Project OCS967P, (1997). *Report on the Interim Benchmarking Exercise: Project Milestone 02/01*, School of Geography, University of Leeds.

Masters, T. (1995). *Neural, Novel & Hybrid Algorithms for Time Series Prediction*, John Wiley & Sons, New York.

Minns, A.W., and Hall, M.J. (1996). Artificial neural networks as rainfall-runoff models, *Hydrological Sciences Journal*, 41, 399-417.

Moore, R.J., Jones, D.A., Black, K.B., Austin, R.M., Carrington, D.S., Tinnion, M. and Akhondi, A. (1994). RFFS and HYRAD: Integrated Systems for Rainfall and River Flow Forecasting in Real-time and their Application in Yorkshire, *BHS Occasional Paper No. 4*, 12.

Openshaw, S., and Openshaw, C. (1997). *Artificial Intelligence in Geography*, John Wiley & Sons, London.

Smith, J., and Eli, R.N. (1995). Neural network models of rainfall-runoff process, *Journal of Water Resources Planning and Management*, 121, 499-509.

SNNS Group, (1990-95). *SNNS User Manual v.4.1*, Institute for Parallel and Distributed High-Performance Systems (IPVR), University of Stuttgart, Stuttgart.

Van der Voort, M., Dougherty, M.S., and Watson S.M. (1996). Combining Kohonen maps with ARIMA time series models to forecast traffic flow. *Transportation Research C*, 4(5), 307-318.

Yakowitz, S.J. (1985). Markov flow models and the flood warning problem, *Water Resources Research*, 21, 81-88.

Yang, R. 1997. *Application of Neural Networks and Genetic Algorithms to Modelling Flood Discharges and Urban Water Quality*. Unpublished PhD thesis, University of Manchester.

Yapo, P., Sorooshian, S., and Gupta, V. (1993). A Markov chain flow model for flood forecasting, *Water Resources Research*, 29, 2427-2436.

Learning Spatial Relationships: Some Approaches

R I McKay & R A Pearson
Computer Science Dept, University College,
Australian Defence Force Academy
rim@cs.adfa.oz.au, +61 6 268 8169
rap@cs.adfa.oz.au

P A Whigham
Division of Land and Water Resources
Commonwealth Scientific & Industrial Research Organisation
paw@cs.adfa.oz.au

*Presented at the second annual conference of GeoComputation '97 & SIRC '97,
University of Otago, New Zealand, 26-29 August 1997*

Abstract

We consider three approaches to learning natural resource models involving spatial relationships, based respectively on decision tree learning, genetic programming and inductive logic programming. In each case, the results of spatial learning on a natural resource problem are compared with the results of non-spatial learning from the same data, and improvements in predictivity or simplicity of the models are noted. We argue also that it is highly desirable that spatial learning systems for natural resource problems incorporate mechanisms for the user specification of learning biases.

1. Introduction

1.1. Machine Learning for Natural Resource Problems

With today's increasing emphasis on environmental limits, the need for accurate and timely information on natural resource issues is pressing. In many cases, the information required for decisions may be expensive to obtain, yet data on some of the underlying variables is relatively inexpensive and available in enormous quantity. The problem is to convert this plentiful data into useful information; machine learning and related data mining techniques provide one promising means to do so.

There have been a number of such applications (for example Barbanente et al 1992; Eklund & Salim 1993; Papp, Dowe and Cox 1993; Stockwell et al 1990; Walker & Cocks 1990). Yet the range is perhaps less than one might expect. Part of the reason lies in the form of the readily available, industrial quality learning systems (Breiman et al 1984; Quinlan 1986). These systems are attribute based, rather than relational - thus they cannot directly learn about spatial relationships. Yet spatial relationships are at the core of many, probably most, natural resource problems.

This paper aims to demonstrate the value of spatial learning, by describing a number of experiments using different methods which have been carried out at University College, ADFA.

Of course, we are not alone in such work. Of recent years, spatial regression methods have appeared in statistical packages (Bowman 1997). However it is well known (Stockwell et al 1990) that discrete machine learning methods outperform regression methods on some datasets. Closer to our approach is the work of (Dibble 1994), which uses an evolutionary approach distantly related to the (Whigham 1996) work reported here.

1.2. Why is Spatial Learning Hard

Spatial problems are intrinsically relational rather than attribute based: they are about the relationships between attributes of particular locations and regions, rather than simply about the local values of those attributes. While particular spatial relationships can often be reduced to spatial attributes (see the discussion below), the reduction requires a-priori knowledge, about the significance of particular spatial relationships for the problem at hand,

which is often not available.

On the other hand, relational learning is intrinsically difficult. The concept spaces to be searched are orders of magnitude larger than those encountered in attribute-based learning.

Furthermore, there are special difficulties with spatial learning problems. Most attribute-based learning, and much relational learning, makes use of greedy search algorithms, which require each new element of the learned model to contribute significantly toward the accuracy of the model. There is no look-ahead: the new element has to make the contribution on its own, without the assistance of any other element. But spatial relationships typically do not make such isolated contributions: they work together with the attributes of the related locations to contribute toward the reliability of the model.

1.3. The Importance of Bias

The machine learning community has gradually come to appreciate the importance of bias in learning systems, and indeed the impossibility of the once-holy grail of unbiased learning (Wolpert and Macready 1995).

In natural resource problems, it is commonly the case that experts in the field have considerable knowledge about the likely forms of models, even if they do not know the exact model at the time.

Taking all this, together with the inherent computational difficulties of spatial learning, it seems clear that systems which provide the user with opportunities to control the bias of the search, and thus reduce the computational cost of the learning process, will be highly desirable for spatial learning in natural resource problems.

2. Sample Problems

Our work to date has been particularly based on two natural resource learning problems. The first is highly atypical, and is specifically chosen because we already know the answer to the problem, and can thus assess sensibly how different learning systems are behaving in relation to that answer. The second was chosen as a fairly typical example of a natural resource problem, and indeed has previously been intensively studied in a purely attribute-based setting (Stockwell et al 1990)

2.1. The Wetness Index Problem

The wetness index problem derives from a pre-existing expert system, LMAS (Whigham and Davis, 1989). LMAS is used to assist with environmental management at Puckapunyal army base in Victoria, Australia. It predicts, from meteorological records and spatial databases describing the site, the likely ground disturbance effects of a given armoured exercise.

One module of LMAS uses the landform and slope layers of the GIS describing Puckapunyal to predict the propensity of particular areas to become waterlogged - the wetness index, with 6 possible values: unknown, dry, average, wet, seasonally waterlogged, waterlogged. This module, like the rest of LMAS, was derived through the traditional expert systems process - as an encoding of the pre-existing knowledge of a geographical expert - and was then validated by ground-truthing. A map of the wetness index for Puckapunyal is given in Figure 1.

The wetness index learning problem is this. The system is given a three-layer dataset consisting of the original landform and slope layers, together with a new layer consisting of the wetness indices as derived by the wetness module of LMAS. The dataset consists of 3,272 polygons, together with a table of the adjacencies between polygons. The system is to learn a new set of rules, which are to predict the wetness index as accurately as possible from the landform and slope layers, together with the adjacency relations.

This particular problem is of interest for three reasons. First, we know that there is a perfectly accurate model of this problem - the wetness module of LMAS. Second, we know that the model involves spatial reasoning, so it is likely that spatial learning will be useful for the problem. Finally, we know the form of the LMAS model, so that if a particular learning system fails to learn well, we can investigate why it does not discover the LMAS solution. On the other hand, the problem is artificial, in that the model we

are attempting to learn is that which best fits the original expert's model of the situation, rather than some underlying "real World" description.

2.2. The Greater Glider Problem

The greater glider dataset is described in detail in (Stockwell et al 1990); briefly, it consists of a 20*20 grid of cells. For each cell, the values of seven independent variables are recorded: the degree of development (D - 3 categories); whether a stream corridor (ST - 2 categories); stand condition from a forestry perspective (SC - 6 categories); site quality from a forestry perspective (SQ - 4 categories); floristic nutrients (FN - 4 categories); slope (S - 3 categories); and erosion (E - 3 categories) (NB in the study area, all sites were highly eroded, E=3, so the erosion attribute may be effectively ignored). For each cell, we also have a value for the putative dependent variable, the greater glider density (GD - 4 categories, ranging from 0-absent to 3-abundant). A map is given in Figure 2.

3. *Simulating Spatial Learning with Attribute-Based Systems*

The first series of experiments described here were performed with the aim of demonstrating that the capacity to learn spatial relations could improve the predictivity of machine learning systems applied to natural resource data. The data used was the greater glider dataset described above.

3.1. Experiments

The experiments were conducted using the Rulefinder decision tree induction system (Pearson 1996). Full details of the experiments and results are given in (Pearson and McKay 1996). Briefly, a first experiment was conducted to provide a baseline for comparison by setting up the conditions as similarly as possible to the experiments of Stockwell et al (1996); a second baseline experiment varied the underlying learning conditions to be similar to those of our main experiments as possible, but without incorporating any spatial information. These experiments led into the main work, in which spatial relationships, built from the underlying attributes, were encoded as additional attributes and added to the dataset.

Taking as an example the underlying attribute "site quality", describing the forestry potential of a location, the relationships encoded as attributes for the various experiments were:

experiment 3: distance to nearest location with a particular value of site quality

experiment 5: whether some adjacent location has a particular site quality

experiment 4: whether there was an adjacency chain of a given length (i.e. A adjacent to B adjacent to C ....) to a location having a particular site quality

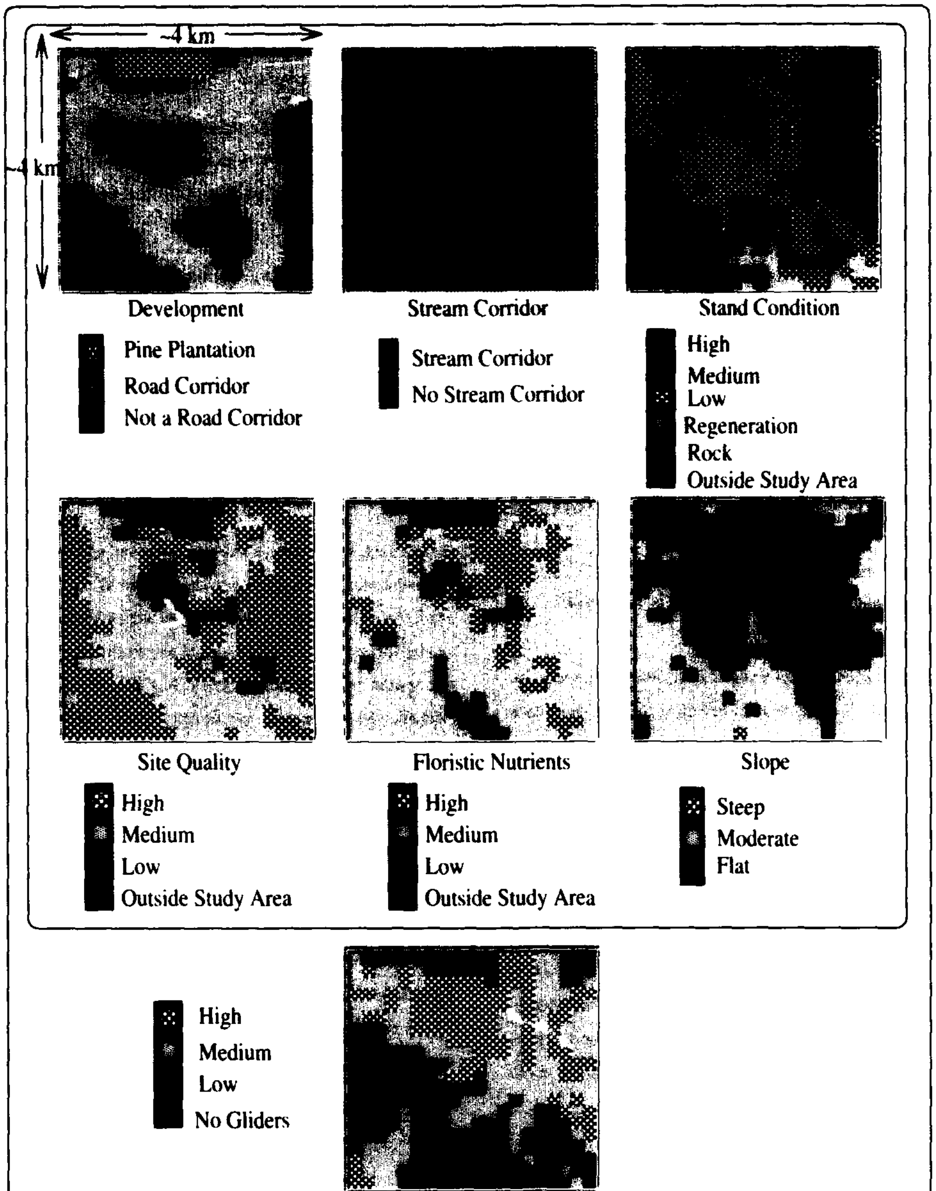
Finally, each of the above experiments was split into two experiments, according to whether values of the learning attribute - the glider density (at sites other than the particular location in question) - were incorporated amongst the spatial relationships encoded (e.g. in experiment 3a, "distance to the nearest site having a glider density of 3" was not encoded as an attribute in the dataset; in experiment 3b, it was so encoded).

3.2. Results

Size and accuracy of decision trees induced from the greater glider dataset.

Experiment	1	2	3a	3b	4a	4b	5a	5b
Tree Size	21	5	27	23	68	13	39	15
Error Rate (%)	47.5	47.75	29.5	29.0	28.75	31.75	34.5	31.75
Std Dev (%)	NA	6.74	5.68	3.71	4.64	7.86	6.78	6.46

Results in the two baseline experiments were very comparable with Stockwell et al (1990), with error rates of 47.5% and 47.75% respectively, and trees of very similar structure. Experiments 3 to 5 gave dramatically improved


~4 km
~4 km
Development
Pine Plantation
Road Corridor
Not a Road Corridor
Stream Corridor
Stream Corridor
No Stream Corridor
Stand Condition
High
Medium
Low
Regeneration
Rock
Outside Study Area
Site Quality
High
Medium
Low
Outside Study Area
Floristic Nutrients
High
Medium
Low
Outside Study Area
Slope
Steep
Moderate
Flat
High
Medium
Low
No Gliders

Glider Den[illegible]

error rates, ranging from 28.75% to 34.5%.

The tenfold cross-validation method, which Rulefinder uses to estimate error rates, also permits the estimation of standard deviation of the error rates. It is thus possible to say that the results in experiments 3 through 5 are significantly different from the results in experiments 1 and 2 (and thus from the Stockwell et al (1996) results) at the 1% confidence level: but they are not significantly different from each other.

One other point to note: the trees learnt here may be approaching the limit of what can be learnt from this data, due to inherent noise and/or missing variables. As shown in Stockwell et al, simply looking at cases in which pairs of cells with the same values for all the independent attributes nevertheless have differing values of the learning attribute, gives an error rate of 24.2%, with a standard deviation of 1.2%. While one should be careful in extrapolating this to spatial learning - since spatial learning in effect provides additional independent attributes by which cells may be distinguished - the similarity of these error rates may not be entirely coincidental.

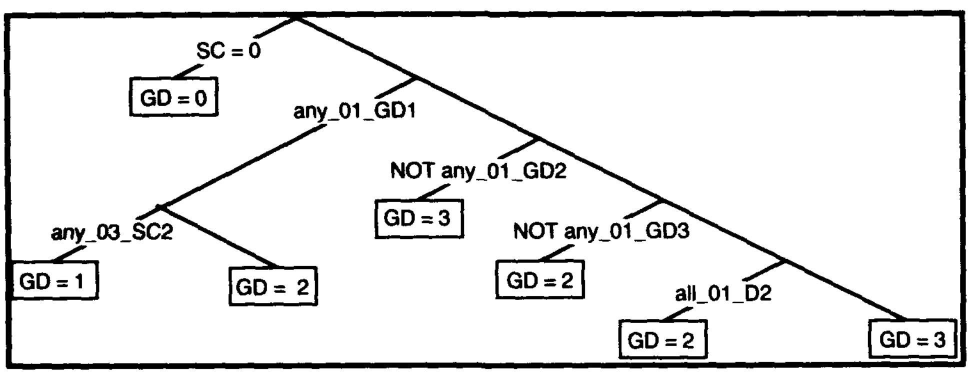
3.3. Discussion

There is always the possibility that the decision trees in experiments 3 to 5 are overfitted to the data. The pruning process in decision tree learning normally provides some protection against this. However the incorporation of spatially derived attributes in the dataset implies that it is not possible any longer to guarantee the independence of the training and test sets, and thus overfitting cannot be ruled out.

However, consideration of the meanings of the decision trees gives some degree of protection against overfitting: on the assumption that the search space of decision trees is sparsely populated with sensible explanatory trees, it is highly likely that any overfitting will be accompanied by meaningless expressions at the tips of the decision trees. Analysis of experiments 3 to 5 suggests that the largest decision trees generated - a 68-node tree in experiment 4a, and possibly a 39-node tree in experiment 5a - may be somewhat overfitted, but that the other treees, which are roughly comparable in size with those of Stockwell et al (1996), are unlikely to be overfitted. A detailed discussion may be found in Pearson & McKay (1996). The smallest tree, that from experiment 4b, is shown in Figure 3.

Thus our final conclusion is that the incorporation of spatial information into a learning process can lead to significant improvements in the predictivity of the models generated. However, the process used is relatively clumsy. It requires the experimenter to know ahead of time which spatial attributes are important, so that they can be incorporated into attributes for use in the learning process. Further, it requires the experimenter to write special-purpose programs to translate the selected spatial relationships into tabular attribute form.

We would naturally prefer that the learning system be able to discover the important spatial relationships for itself,



while permitting the user to narrow the focus of the learning to particular classes of spatial - or other - relationships if such knowledge is available. Thus a prime focus of our work has been on learning systems which can work directly with spatial relationships, but permit the user to vary the bias of the learning space search.

4. Genetic Programming and Geospatial Relations

The work on context free grammars for genetic programming (CFG-GP) discussed here is reported in detail in the doctoral thesis of P A Whigham (1996). It builds upon the genetic programming paradigm of Koza (1992). However, in the genetic programming paradigm, the description language is a by-product of the GP system and is not amenable to user variation except through re-building the underlying system.

In line with our conviction that useful geospatial learning systems will require simple mechanisms by which the user may specify the search space the learning system is to use, CFG-GP provides a context-free grammar in which the user defines a grammar for the language the learning system is to use for the specific problem (this work follows on from the Grendel system (Cohen 1994), which used context free grammars similarly, but within the inductive logic programming paradigm).

The greater glider dataset contains a number of hard constraints. For example, a small proportion of the cells are rated as "outside the study area". These cells have their glider density set arbitrarily to zero. This causes little problem to deterministic learning systems such as decision tree systems: these rapidly learn that "outside the study area" implies "glider density zero", and are thus free to ignore those cells from that point on (indeed, this is the top-level decision in virtually all the decision trees we have generated from this data).

A stochastic learning paradigm has problems with such hard constraints, since the system will always be prepared, even though with low probability, to re-visit these constraints and to try alternatives. Whatever mechanism is used to evaluate the success of the system will thus incorporate some penalty for this willingness to try alternatives.

Fortunately, CFG-GP incorporates a mechanism for investigating this effect. The user may explicitly incorporate the hard constraint into the search language used by the system, so that the option of revisiting the constraint is no longer available.

4.1. Experiments

CFG-GP was first applied to the greater glider dataset in non-spatial mode. A number of experiments were conducted, starting off with a simple attribute language describing the dataset, then extending this with two hard constraints: the "outside search area" constraint described above, and a second explicitly requiring the system to learn descriptions for each of the four glider density classes (otherwise the system may simply ignore density classes which are sparsely represented in the data).

The language was then extended with additional spatial expressions. For each possible value V of each of the underlying attributes A, and for each distance D, the system is permitted to derive the boolean expression determining whether there is a cell within distance D of the current cell, in which the attribute A has the value V.

For computational reasons (genetic programming is computationally very expensive), the values of D were limited to be either 1 or 2, though the decision tree work above suggests that distance values up to 5 may be meaningful in this dataset.

4.2. Results

In the simplest attribute learning example above, the system achieved an error rate of 47.5 ± 3.4% (based on 6 trials). Incorporating the hard constraints mentioned above improved the learning somewhat, to an error rate of 42.9 ± 3.2% (6 trials). Finally, addition of spatial expressions gave error rates of 32.8 ± 1.7% (6 trials). The best ruleset was:

if ((stand_condition = rock)
  or ((slope > flat within distance 2)
    and (stand_condition = regeneration within distance 4)
    and (floristic_nutrients > medium within distance 5)))
then glider_density = low
else **if** ((slope > flat within distance 2)
    or (stand_condition = regeneration within distance 4))
  then glider_density = medium
  else glider_density = high

4.3. Discussion

In non-spatial learning, CFG-GP achieved similar results to Stockwell et al (1990), and to the Rulefinder results reported above (the incorporation of hard constraints improved the learning, but the improvements are only marginally significant). Significant improvements were obtained by the incorporation of spatial information into the learning; the improvements are very comparable with those achieved by Rulefinder, providing further confirmation that the improvements in error rate are real, and not just the result of overfitting the data.

5. *Inductive Logic Programming and Geospatial Relations*

We have previously (McKay 1994) reported negative results in the application of ILP systems to geospatial learning problems. Our analysis there pointed out that the lack of results were not due to inherent limitations of the ILP paradigm, but were particularly related to specific assumptions made in the greedy algorithms used.

Specifically, the systems assumed that useful relationships either directly reduce dataset noise (without the assistance of subsidiary attributes), or are determinate. Unfortunately, spatial relationships such as distance, relative orientation etc. do not have either of these properties, so that spatial relationships would never be tested by these algorithms.

Since that time, we have carried out further experiments with the more recent Progol system (Muggleton 1995), which does not make determinacy assumptions. Progol learns logical rules, in the form of prolog programs. Progol does not handle noise well, so we have not gained any useful results in learning from the greater glider dataset. However experiments with the wetness index dataset have yielded some interesting results

5.1. Experiments

In the first experiment, progol was run on the wetness index as described above. The second experiment was identical, except that the table of adjacencies was deleted from the dataset, so that progol could only learn attribute descriptions of the dataset.

5.2. Results

Progol learns a complete description of the dataset on which it is run. If necessary, it will generate rules for the dataset cell by cell, in order to do so. Unlike Rulefinder and CFG-GP, it does not provide for a separation of learning and test datasets. Thus results from Progol do not give meaningful error estimates. The only meaningful comparison we can make is between the sizes of the rulesets learnt in each run. Note also, that these rules have been learned from positive data only: since progol was unable to deduce that "dry" and "average" are incompatable, it was prepared to learn identical rules for both. Further work, to ameliorate this problem, is in progress.

The first run, incorporating adjacencies, described the dataset with 6 rules, using 20 conditions (note that the land unit types are ordered):

wi(A,wet) **if** land_unit(A,B) **and**
  B > *floodplain_seasonally_inundated*
wi(A,dry) **if** land_unit(A,B) **and**
  B < dam **and** B > floodplain_seasonally_inundated
wi(A,average) **if** land_unit(A,B) **and**
  B < dam **and** B > floodplain_seasonally_inundated
wi(A,wet) **if** A adjacent_to B **and**
  slope(B,C **and** C > -3
wi(A,seasonally_waterlogged) **if** slope(A,B) **and**
  A adjacent_to C **and** land_unit(C,D) **and**
  D < sand_dunes **and** D > floodplain_seasonally_inundated
wi(A,waterlogged) **if** A adjacent_to B **and** B adjacent_to C **and**
  slope(C,D) **and** D > -2.

The second run, omitting adjacencies, required 10 rules and 40 conditions.

The original expert ruleset, when expressed in the above language, has 13 rules and 42 literals.

5.3. Analysis

The most important result is that experiment 1, using spatial learning, learnt a very much simpler model of the dataset than experiment 2, using purely attribute learning. The big difference lies in only one of the wetness index values: in experiment 1, "wet" cells are described in one spatial and one non-spatial rule, using 8 literals. In experiment 2, 5 non-spatial rules are required, using 25 literals.

Secondly, it is interesting that progol has learnt a model which is simpler, in this language, than the original expert ruleset. The comparison is not entirely fair, however: the expert ruleset was originally expressed in a completely different language, and its present size is partly a result of the translation process. Moreover, the expert ruleset did know about such issues as mutual exclusiveness of wetness values. Nevertheless, it is fair to say that the spatial learning process has produced a ruleset which is smaller and simpler than the non-spatial process, and of expert quality in these respects.

6. Conclusions

Learning systems which can take spatial relationships into account may learn more accurate models than non-spatial learning systems, in real-World natural resource problems. The genetic programming and inductive logic programming paradigms both provide mechanisms with which to attack such problems. So far, greater success has been achieved with GP approaches than with ILP, but this does not seem to be due to any inherent limitations of ILP. Assuming that ILP systems able to handle both noise and indeterminacy become available, the choice between the two may come down to ease of use vs computational complexity: correctly setting up an ILP system may require greater understanding than an equivalent GP system, but the GP system is likely to use more computational resources. As an indication, the CFG-GP work reported above required cpu-days on a SUN SPARC 1000. ILP is also computationally expensive, but more on a scale of cpu-hours than cpu-days.

All existing relational learning systems are computationally expensive; this is unlikely to change, as relational learning is an inherently difficult task. But experts working with geospatial datasets typically have considerable knowledge about constraints on the likely structure of models of those datasets - often arising from knowledge about the physical and other processes involved. Thus it is highly desirable that learning systems for use in geospatial problems permit the user to incorporate this knowledge in the search strategy of the learning system involved. The Grendel and CFG-GP systems mentioned above (along with many other learning systems) give indications of how this may be achieved. A useful by-product of the use of such biases is the possibility of assembling a body of knowledge about useful biases for geospatial learning, and thus of the overall structure of spatial knowledge.

7. References

Barbanente, A, D Borri, F Esposito, P Leo, G *Maciocco and* F Selicato (1992) Automatically Acquiring Knowledge by Digital Maps in Artificial Intelligence Planning Techniques. Theories and Methods of Spatio-Temporal Reasoning in Geographic Space. A U Frank, I Campari and U Formentini, editors. *Springer Lecture Notes in Computer Science* 639. Springer Verlag, Berlin, Germany, 1992, pp 379 - 401.

Bowman, a (1997) S+ SpatialStats: A Spatial Statistics module for S-Plus. *Maths&Stats Newsletter*. CTI Statistics, University of Glasgow, Glasgow, Scotland, 1997.

Breiman, L, J H Friedman, R A Olshen and C J Stone (1984) Classification and Regression Trees. Wadsworth Inc, Belmont, USA, 1984.

Cohen, W W (1994) Grammatically Biased Learning: Learning Logic Programs Using an Explicit Antecedent Description Language. *Artificial Intelligence* 68(2), 1994, pp 303-366.

Dibble, C (1994) Beyond Data: Handling Spatial and Analytical Contexts with Genetics Based Machine Learning. *Sixth International Symposium on Spatial Data Handling* Edinburgh, Scotland, September 1994

Eklund, P W and A Salim (1993) An Experiment with Automated Acquisition of Classification Heuristics *Conference on Advanced Remote Sensing* University of NSW, Sydney, Australia, 1993, volume 2, pp 83 - 94.

Koza, J R (1992) Genetic Programming: On the Programming of Computers by Means of Natural Selection. Bradford, MIT Press, Cambridge, USA, 1992.

McKay, R I (1994) Relational Learning for Geospatial Problems. *ICARCV Workshop on Spatial and Temporal Interaction: Representation and Reasoning*, Singapore, 1994.

Muggleton, S (1995) Inverse Entailment and Progol. *New Generation Computing Journal* 13, 1995, pp 245 - 286.

Papp, E, D L Dowe and S J D Cox (1993) Spectral Classification of Radiometric Data using an Information Theory Approach *Conference on Advanced Remote Sensing* University of NSW, Sydney, Australia, 1993, volume 2, pp 223 - 232.

Pearson, R A (1996) Single Pass Constructive Induction with Continuous Variables. *ISIS: Information, Statistics and Induction in Science*. D L Dowe, K B Korb and J A Oliver, editors. World Scientific, Singapore, 1996, pp 31 - 42.

Pearson, R A and R I McKay (1996) Spatial Induction for Natural Resource Problems: A Case Study in Wildlife Density Prediction. Technical Report CS11/96, School of Computer Science, University College, University of New South Wales, Canberra, Australia, 1996. Submitted to *AI Applications in Natural Resource Management*.

Quinlan, J R (1986) Induction of Decision Trees. *Machine Learning* 1, 1986, pp 81 - 106.

Stockwell, D R B, S M Davey, J R Davis and I R Noble (1990) Using Induction of Decision Trees to Predict Greater Glider Density. *A I Applications in Natural Resource Management* 4(4), 1990, pp 33 - 43.

Walker, P A and K D Cocks (1990) Habitat: a Procedure for Modelling a Disjoint Environmental Envelope for a Plant or Animal Species. *Global Ecology and Biogeography Letters* 1, 1990, pp 448 - 461.

Whigham, P A (1996) Grammatical Bias for Evolutionary Learning. PhD Thesis, University College, University of New South Wales, Canberra, Australia, 1996.

Whigham, P A and J R Davis (1989) Modelling with an Integrated GIS/Expert System *Ninth Annual ESRI Users Conference*, Palm Spring, ESRI, Redlands, USA, 1989.

Wolpert, D H and W G Macready (1995) No Free Lunch Theorems for Search. Santa Fe Working Paper, Santa Fe Institute, Santa Fe, USA, 1995.

8. Acknowledgements

We wish to thank Dr P Laut, Dr R Davis and Ms S Cuddy for use of the wetness index dataset, and Dr S Davey and Dr D Stockwell for use of the greater glider dataset.

Conceptual data modelling in an archaeological GIS

Paul Rivett
Anthropology Department
University of Otago
Dunedin

Presented at the second annual conference of GeoComputation '97 & SIRC '97, University of Otago, New Zealand, 26-29 August 1997

Abstract

Recent discussion of archaeological GIS method and theory has centred around a debate concerning the use of the technology. This paper argues that key problems in this debate can be overcome by looking at how data are defined and structured with regards to the overall project. It specifically deals with two points. First, that an appropriate theoretical framework needs to be developed and that this should occur at the level of the data. Second, recent debate has overlooked the importance of database design and data structure at the conceptual level. Conceptual data models provide a link between reality as it is perceived by humans and the way in which reality will be represented in the database. A spatially extended entity relationship (SEER) conceptual data model is developed for an archaeological GIS which will make explicit any relationships (both spatial and non-spatial). A hermeneutic methodology is outlined that will ensure that the conceptual model developed will accurately reflect the dynamic nature of the data. The data itself comes from a case study on the distribution of archaeological sites in Northeast Thailand.

1. Introduction

Although geographical information systems (GIS) can no longer be regarded as a new technology, within much of the archaeological literature attempts are still being made to 'show the usefulness of GIS in archaeology'. These studies repeat things that have been said many times in the past. It can now be stated with some confidence that we know that GIS are useful - the time has come to develop an appropriate theoretical basis for the use of the technology within archaeology. This is an area that has not been addressed and there is an ongoing concern about "the general lack of an underlying theoretical basis for understanding spatial and temporal data within the context of a given discipline" (Burrough and Frank 1995:102).

This suggestion that GIS are *discipline-independent* has important implications for archaeology. No longer should we wait for developments in associated disciplines, it has become critical that we develop an appropriate theoretical framework from which to utilise GIS. Discussions regarding this point have been intimated in the archaeological literature, but they lean towards more general discussions concerning the future directions of the technology (e.g. see Limp 1996). In many cases, there appears to be a kind of 'technological determinism' involved with the technology itself directing the nature of the research rather than the research being the primary focus and the technology a the tool.

This paper argues that theoretical developn. ts should occur at the level of the data, not at the level of the technology. Technological advancements can only lead us so far; although there are many areas which need redressing, it is stressed here that for the development of appropriate method and theory we must turn to the fundamental level of a GIS - the database. The database is essentially the foundation from which the system is built and without this there would be no GIS. Data modelling and database struc-

ture are issues that have not been addressed in the archaeological literature which is worrying as it is here that data and their various relationships are defined. The following sections discuss the current use of GIS in archaeology; from this, several problems and limitations are identified in the use of GIS and it is argued that these problems can be traced back to the fact that there is no consistent theoretical framework from which to utilise GIS. A hermeneutic methodology is outlined that will ensure explicit data definition for the maintenance of data structure and data integrity. This methodology is discussed with an archaeological example from Northeast Thailand.

2. GIS in archaeology

This section shows the need for the development of a consistent theoretical base for the utilisation of GIS within archaeology. It discusses the previous uses of the technology and outlines the need for a substantive geographical information theory.

Harris and Lock (1990, 1995) and Kvamme (1989, 1995) have discussed the history of GIS in Europe and North America respectively and they note a fundamental difference in the use of GIS between the two. This is most evident when applications are compared between these two areas (for Europe see Lock and Stancic 1995 and Bietti *et al.* 1996; for North America see Allen *et al.* 1990 and Aldenderfer and Maschner 1996). First, and generally within a North American context, emphasis is placed on a functionalist, or processual, approach to explanation. It is argued that human behaviour is non-random and that general patterns can be seen in the archaeological record. These patterns are created by people interacting with the natural environment and can be identified in a statistically meaningful way. This allows for mathematical formulations to be developed that allow for the prediction of sites and simulation modelling. This approach treats space in a Cartesian manner largely devoid of social meaning. Second, other writers, predominantly from Europe, argue that space is socially produced and its manifestation on the landscape depends on its particular context and cannot be generalised. They argue that human behaviour is unpredictable and that patterns seen in the archaeological record can be misleading (Hodder 1982). They are interested in what they call the 'social landscape' and include attempts at 'rehumanising' GIS.

2.1 The environmental modelling approach

The use of GIS in the first instance developed from archaeologists interested in examining the relationship between archaeological sites and various environmental conditions. These associations were statistically defined and this facilitates the development of models from which to predict site location within a given area. For such purposes GIS is an excellent tool, but it must be acknowledged that there is no explanatory power in this method (Voorrips 1996). In fact, the use of GIS in this manner can lead to the exposition of an outdated environmentally deterministic argument. For example, Brandt *et al.* (1992) develop a model for the prediction of site location in the Netherlands. Due to harsh vegetation and alluvium deposits surface surveys are difficult to undertake; therefore, the development of a predictive model would facilitate site recovery. They note that environmental data are being used as they are "easy to obtain for a region" (Brandt *et al.* 1992:269) and since social variables must be reconstructed for each period, which is "a task often beyond our data retrieval possibilities" (Brandt *et al.* 1992:269), they do not incorporate such data into the analysis. They further restrict their study by limiting themselves to "simple associations between sites and modern map categories" (Brandt *et al.* 1992:272). Such restrictions mean that they cannot say anything useful about prehistoric behaviour, and although behaviour could be inferred from such relationships, their lack of interest in social variables rules out inferences of this type. In a more explanatory approach, Hunt (1992) undertook the analysis of site catchments in the Late Woodland Period (A.D.1000-1600) in Western New York State. The catchment area is "the zone of resources, both wild and domestic, that occur within reasonable walking distance of a given village" (Flannery 1976:91). The GIS was used to determine soil productivity in each of the site catchments and it was concluded

that villages were established in areas suitable for the production of maize. Again this study is concerned solely with environmental and not cultural data. Although it is not necessarily an environmentally deterministic approach, the relationships that are developed are obvious and one does not need a GIS for their confirmation.

2.2 The 'social landscape' approach

Strong criticisms of the situation of GIS within such an explanatory framework came from various researchers whose theoretical orientations are sympathetic to the second group. Wheatley (1993:133) stresses the need to move away from such functional interpretations as they are "an extremely restrictive approach to archaeological explanation." Furthermore, Gaffney *et al.* (1995:211) note that:

> "there are good reasons to suggest that the application of GIS techniques in such a way could ultimately prove to be restrictive to the general development of archaeological thought. In its least harmful form, the indiscriminant use of GIS solely in conjunction with mapped physical data may result in the slick, but repetitious, confirmation of otherwise obvious relationships. In the worst case, it might involve the unwitting exposition of an environmentally or functionally determinist analytical viewpoint of a type which has largely been rejected by the archaeological community."

What these and other authors suggest is the need for the incorporation into a GIS of theory laden data representative of a culture. This type of argument is firmly linked to the 'post-processual' tradition of thought stemming from England where there has been an increasing interest in the social production of space and its physical and temporal manifestation across a landscape (Bender 1993; Thomas 1993; Tilley 1994). This theoretical awareness has been alluded to in many GIS studies and this has been a necessary development to move away from the limited explanation under a processual approach.

The main approach so far used for 'rehumanising' GIS has been viewshed analysis (Wheatley 1995). It is argued that this method provides a means for incorporating human perception into a GIS analysis. For example, viewshed analysis has been applied for the determination of visibility between monuments over a landscape. Wheatley (1995) provides an analysis of the intervisibility of long barrows in two separate areas of Neolithic Britain and shows that between these areas there is a difference in visibility. From this a post-processual interpretation is offered concerning the control of the monuments enabling the legitimacy and perpetuation of ones own status and authority through the historic importance of the extant monuments. This analysis has several limitations. First, it uses limited data sets; for example, only topography and the location of the monuments are used. No consideration is given to any other variables, be they other sites or even other basic environmental variables. Second, the actual study uses the ground surface as the basis for inferring line of sight; this does not necessarily suggest intervisibility as it was noted that the prehistoric vegetation was considerably greater than at present. Finally, such studies "critically confuse the concept of 'vision' with that of 'perception'" (Gillings 1996:79). Just because two monuments are intervisible, or visible from various parts of the landscape, does not necessarily imply a relationship of importance to the prehistoric individual. Here they distinguish between perception as sensation and perception as cognition (Rodaway 1994). There is a continual dialogical relationship between simple acts of vision (sensation) and mental process (cognition) which enable the individual to create a geographical understanding - a sense of the world. In the archaeological studies using GIS in the realms of 'perception', they have situated the analysis firmly with regard to perception as vision, and have disregarded cognitive aspects which underlie phenomenological approaches to the environment (Tilley 1994).

What these studies show is a supposed change in focus from environmental to cultural concerns. Whereas the former studies are explicit in the use of the environment as a major factor in their analyses, the latter try to downplay the importance of such variables. Although they appear to incorporate cultural variables, in actual fact they provide nothing more than studies based solely on environmental

data - and in this regard, in many instances, there is a reduction of data. The cultural variables stated as part of the analyses are never explicitly defined and regard is only given to them in interpretation.

2.3 Towards a substantive archaeological geographic information theory

The need for the development of a theoretical basis for GIS in archaeology has come after similar discussions of this type in geography and information science. There have been arguments for the development of a geographic information theory dealing with the representation of knowledge (Molenaar 1989), a geographical information science[1] which sees the need to develop generic questions to create a 'core discipline' (Goodchild 1992) and a more holistic 'discipline-independent' theoretically informed post-modern 'theory of spatial relationships' which is both mathematically elegant and in tune with concepts developed in the minds of humans (Burrough and Frank 1995; Mark and Frank 1996). Although these are opposing ideas for the development of a GIS epistemology, they all make the same general point concerning the lack of an underlying theoretical framework - be this as a discipline in itself or as something that must be created for each discipline in its own right. Although each subject area utilising GIS has some inherent spatial component, there are fundamental differences regarding the nature of space; because of this "problems that are specific to the application of GIS in a particular field clearly need to be addressed in the context of that field, and with the benefit of its expertise" (Goodchild 1992:41).

The area of concern in this paper from an archaeological viewpoint is the modelling of data at the conceptual level (Batini *et al.* 1992). There are generally considered to be three levels of abstraction relevant to geographical databases (conceptual model Ë logical model Ë physical model). Conceptual data modelling is the first of these levels. It formalises human concepts of space and is necessary because computer systems work through sets of formal rules. Furthermore, conceptual models are an abstraction of the real world and incorporate only relevant data (Maguire and Dangermond 1991). The other two levels (logical and physical) are to do with implementation issues and storage requirements respectively. For the present purposes, the conceptual model can stand on its own without regard for implementation since, at this stage, we are concerned with explicit data definition rather than the implementation of a database. This will be achieved through the use of hermeneutics (see below); although hermeneutics has been previously used in GIS design (Gould 1994), the concern was with the interaction between the designer and the user. In the case here, hermeneutics is concerned with the interaction and interpretation of data.

The development of appropriate conceptual schemas help to, first, incorporate non-environmental data in order to augment the more common environmental variables within a data model and second, to extend the data model to incorporate abstract semantic mechanisms for the definition of spatial and topological relationships. In the past, the conceptual design of standard relational databases have not accommodated semantics that explicitly define such relationships. Recently, several models have been developed to extend the capabilities of the conceptual schemas in this direction (Fernández and Rusinkiewicz 1993; Firns 1994). Archaeologists have not used traditional entity-relationship (ER) data modelling techniques for the establishment of GIS databases and it seems appropriate, in the light of the preceding discussion that such techniques be employed for due consideration of the data.

The critique above concerning the uses of GIS highlight several basic problems. Concerns regarding the functionalist use of GIS led to the expounding of approaches within the realm of a humanist archaeology. This rehumanising has merely shifted the environmental emphasis to a more subtle position which has narrowed the scope of GIS through the use of limited data sets. The following section

[1] This has been seen in the name change of the International Journal of Geographical Information *Systems* to the International Journal of Geographical Information *Science* (Fisher 1997).

outlines a methodological approach and example for the development of conceptual database schemas essential for the incorporation of explicit in: ›r› ›tion concerning the nature of the data.

3. Using hermeneutics to design an archaeological database

3.1 Hermeneutics

Hermeneutics grew out of attempts to develop a theory of interpretation. Initially it set out to equate social sciences with natural sciences thereby seeing both as following an objective approach to understanding. Gadamer (1975[1960]) reacted against the use of hermeneutics in this manner; rather, he developed the notion of 'prejudice' from Heideggers' 'pre-understanding'. He argues that prejudice and understanding are thoroughly conditioned by the past, a past he calls 'effective history' (Gadamer 1975[1960]:267). Furthermore, the "really central question of hermeneutics" is that of separating "the true prejudices, by which we understand, from the false ones by which we misunderstand" (Gadamer 1975[1960]:266). Although this notion has been critiqued by Habermas (Warnke 1987) and Ricoeur (1981) it is believed to be a useful concept and is used here. Prejudice in the case of data structure for an archaeological GIS is likewise determined by our 'effective history', in this case 'effective knowledge'. Determining the data to be incorporated within the GIS database necessarily involves questioning the assumptions of the analysis and the assumptions the researcher has concerning the study.

Figure 1 outlines the hermeneutic procedure for this study; it identifies prejudice, problem definition and data definition as being major components. However, these three components are not mutually exclusive, rather, there is a continuing dynamism between them. Although it appears an iterative approach, the dynamics involved preclude the definition of a step by step procedure. Interpretation proceeds differently for each individual as it is part of their 'effective knowledge'. Past experience determines the prejudiced notions an individual has; from an individuals' knowledge base, the identification of problems occurring in our understanding of a discipline takes place, and in turn data are defined in order to consider these problems. Such data comes from a variety of sources, and its definition

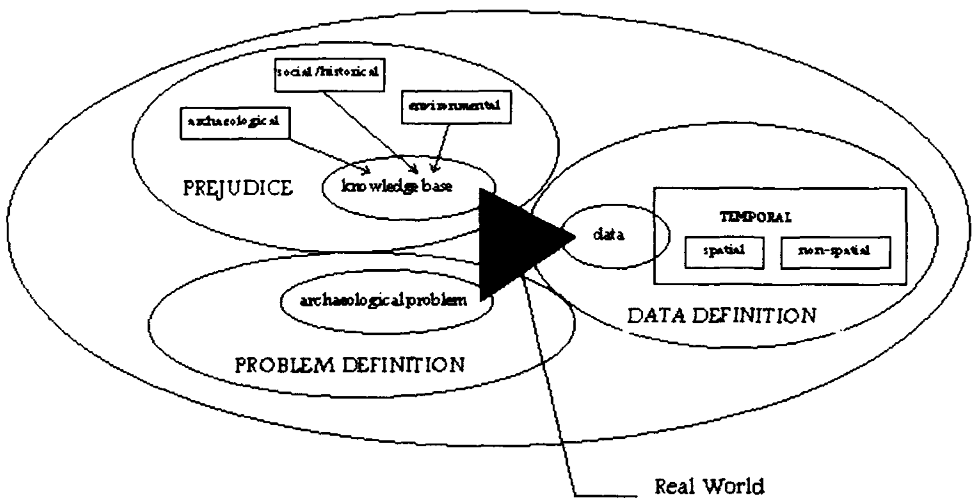


Figure 1: Hermeneutic method for conceptual data modelling for an archaeological GIS

necessitates change in our knowledge base as there is an extension of knowledge and the potential for new problems and different ideas to be developed as a result of data definition. In turn, this information is augmented by input from the 'real world' which is accumulated knowledge not specific to the problem but which more or less alters an individuals' perspective; in a sense, random knowledge. Data come in two forms, spatial (e.g. the distribution of settlements) and non-spatial (data based on ethnographical, anthropological and archaeological sources). These are accumulated and evaluated within the terms of the research and are embedded within an overall temporal framework. At no time do these data have any fixed meaning; this is because "Data do not just 'exist', they have to be created, and who does the creating, for whom, and for what purposes, is vital" (Taylor and Overton 1991:1088).

How then does this translate into a method for generating suitable schema for design of a database? In the preceding discussion we have noted the fact that explicit incorporation of data into an archaeological database does not currently occur at a satisfactory level. The hermeneutic method outlined here necessitates the elaboration of necessary data and does so in an explicit manner.

3.2 The archaeological problem

There have been numerous settlement pattern studies undertaken in Thailand and they are generally concerned with site distribution on a regional level, specifically, the relationship between the distribution of sites and the environment (Higham *et al* 1982; Ho 1992; Moore 1988a; Mudah 1995; Welch and McNeill 1991; Wilen 1987). Previous studies of settlement patterns of a particular type of site in this area, the moated site, has seen the explication of models concerning their development and distribution (Moore 1988a; Welch 1985). The basic model is: settlements were first established on the alluvial plain of the Upper Mun Valley during the Tamyae phase (1000-600B.C.). Intensive forms of agriculture were adopted during the Prasat phase (600-200B.C.) which made possible the expansion of settlement from the alluvial plain to the terrace and upland zones. Expansion to these areas saw the exploitation of salt, iron and timber resources which became important trade items. Welch (1985) was interested in documenting the role of centralisation, urbanisation and agricultural intensification with regard to these sites and their roles in long distance exchange. Moore (1988a) was interested in the moated sites as a technological group and attempted to document their overall structure and distribution. She studied them in isolation from an overall settlement pattern that included the larger moated sites, as well as smaller unmoated sites and rectangular water storage reservoirs (*barays*).

These models overlook a large body of data regarding human societies. Specific community level behaviour cannot be enlightened by such regional analyses. A major assumption in this analysis is that there is some kind of community structure based on the individual site (Trigger 1978). It is not that this structure has so far proven to be elusive to researchers in this area, it is just that it has not been an area in which major research has been undertaken. In order to locate these communities, relationships need to be identified between various factors considered useful for their identification. If the community concept can be identified, a fundamental aspect is the change in such organisation from prehistoric times into the historic Khmer period; a temporal shift of approximately 2500 years. This period saw fundamental changes in religion, symbolism, ideology, technology and social organisation which reached its peak during the time of the Angkorian *mandala* (8-14th centuries A.D.). Although these developments are manifested in monumental structures such as the Khmer temples of Cambodia and Northeastern Thailand, changes in basic community structuring are still largely unknown. We will undoubtedly have to wait until a larger proportion of sites has been excavated, but we can begin by examining spatial relationships of community structure. The archaeological problem is, therefore, to identify these communities both spatially and temporally; this is a problem that GIS can help solve.

3.3 The GIS solution

The archaeological problem identified above is just one

part of the hermeneutic procedure (problem definition). The data contained in the spatially extended entity relationship (SEER) (see Firns 1994) diagram in Figure 2 is the other part (data definition). The critique above questioned the level to which data are defined; the following discussion concerns the SEER model and what it means in terms of this study. The data can be placed into several categories, or locational reference points: soil, hydrology, prehistoric vegetation, sites and temporal period. It will be seen that the first four of the categories are contingent upon temporal period. Each locational reference point is related back to a location which has a specific x, y coordinate value (see figure 2). At this stage we are not concerned with implementation of a database, so such abstraction is useful. These locational references are discussed below.

3.3.1 Hydrology

This category holds information concerning rivers, reservoirs (*barays*) and moats. Since prehistoric times, rivers have moved across the landscape, either naturally as they become flooded or due to human diversion (Welch 1985:292-3). From this, determining the contemporaneity between sites and rivers is very important and although river channels can be dated (Bishop *et al.* 1994), there is no information from this part of Thailand. Instead, we must work by association and relative chronology. The existence of the moated sites and *barays* help in this situation. The function of the moats surrounding these sites are not yet known, but most writers agree on them being used for some kind of water reticulation necessitated by the extremely arid conditions. It is assumed that the moated sites had a water source, and as can be seen on aerial photographs, rivers provide this source. Site abandonment is often linked to the movement of this water source, so as the river moves, so does the site, and the latter can be dated. Furthermore, the *barays*, which are large rectangular storage structures constructed by the Khmer between the 7-14th century AD for domestic, agricultural and religious purposes, were supplied by river and stream diversion (Moore 1988b). So it can be seen that the rivers did not just exist as a natural phenomenon, they played a large part in society. In fact it can be seen that "in no small sense, South East Asia is a region where water - not land - is the defining element and where human-water relationships, not human-land relationships are determining" (Rigg

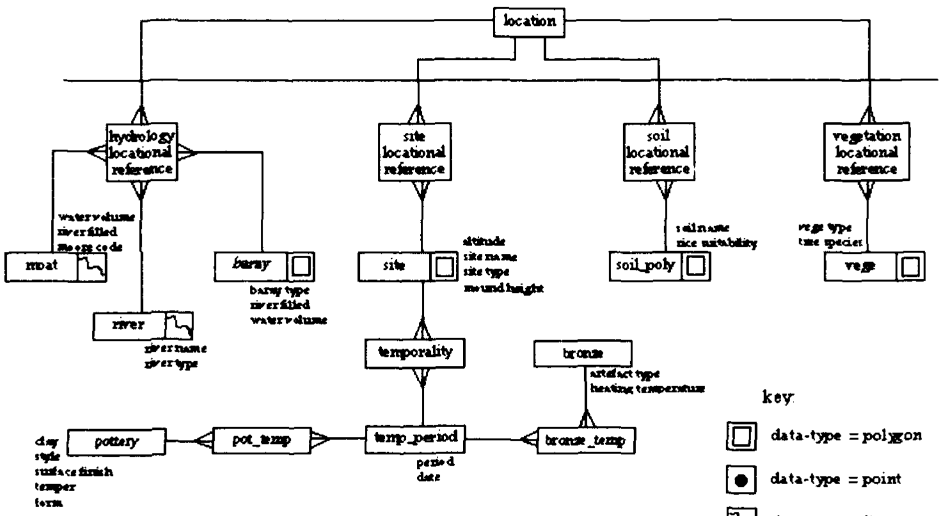


Figure 2. SEER diagram of the archaeological database

1992b:1; also, see papers in Rigg 1992a). An important example for this is the *bun bang fai*, the skyrocket festival of Northeast Thailand where a rocket representing a giant phallus is erected and shot into the sky to fertilise the heavens and to supply rain serving "to remind the male rain god to pour out his semen onto mother earth" (Demaine 1978:52).

3.3.2 Prehistoric Vegetation

This section of the database holds information concerning the prehistoric vegetation of the area. Stott (1978b:7) quotes the 17th century French naturalist and explorer, Nicolas Gervaise, who said that the forests are "so great that they take up more than half of the land...so thick that it is almost impossible to pass through." The vegetation cover of today in no way reflects that of prehistory or indeed of the time of Gervaise. In 1942, 42% of Northeast Thailand was covered in forest, there is now less than 10% (Moore 1992). However, we can recreate the past environment with a good degree of accuracy (van Liere 1980). Deforestation occurred in prehistory although not at a level which seriously altered the nature of the vegetation. Due to the methods of rice cultivation where areas were cleared to increase productivity, soil generally deterioratef and became incapable of supporting any form of plant life other than coarse grass and scrub (Ng 1978). Indeed, over time as more land was cleared such problems undoubtedly increased.

3.3.3 Soil

Basic characteristics regarding soil types are held in this section of the database. Most importantly is the definition of soil suitable for rice cultivation. However, Bayard (1992) has noted several limitations in using soil type as a factor in determining site location and the suitability for rice cultivation. Undoubtedly soil type was important, but it has been exaggerated as a factor in prehistory.

3.3.4 Site

Data regarding the site are important as it is assumed that this is where community structure is to be located. This part of the database holds information concerning basic data on the site; including mound height, size of the mound and the site type (e.g. moated, unmoated, rectangular, territorial and salt making [Moore 1988a]). Each of these types had different functions and can be dated to different periods. Therefore, it is important to define explicitly each type chronologically and once this is done, relationships between the various types can be discerned. Of these types, the salt-making sites are the most ambiguous. Although there are hundreds of mounds scattered throughout Northeast Thailand very few have been excavated (Higham 1977; Nitta 1992). These were important manufacturing sites as salt became a powerful trade item and was used for the preservation of food for consumption during the dry season (Nitta 1992). Salt-making was a dry season activity (Higham 1996a:315) and undoubtedly played a large part in community life. Thus these salt-making sites are essential components in the identification of the community.

As many sites were continuously occupied over long periods and fit into different settlement systems throughout the term of their occupation, there needs to be strict temporal control over their distribution at certain times in prehistory. This entity is linked to temporal period for this purpose.

3.3.5 Temporal Period

This aspect of the database is the most important as it is here where non-environmental variables are defined. Variables such as language (Higham 1996a, 1996b), religion (Tambiah 1970), burial practices (Higham *et al.* 1992), trade (Glover *et al.* 1992), along with information regarding bronze (Pigott *et al.* 1992), iron (Pornchai 1992) and pottery (Bayard 1977) technologies allow communities to be located at a given temporal period. Possibly the most useful indicator of temporal period is pottery which is an artefact type that has huge diversity in form, uses and style. These aspects along with rim-form, surface finish, surface-texture, colour, and temper help to differentiate between pottery types of different periods, but it is the general attributes of particular styles that are important rather than specific aspects. Therefore, incorporation of such variables could be considered useful for a regional/nationwide database,

but for the present study they are not deemed necessary; strict associations between pottery types and temporal period will suffice.

The most important relationship is between this entity and the site entity. The site entity holds only that information for the physical nature of the site whereas the temporal period entity holds data that defines the activities at a site at a particular time. It is these activities that allow temporal relationships between the various entity sets to be defined. Furthermore, it is important to note that although the locational references discussed are environmental, they are embedded within a social context making it extremely difficult to make general conclusions regarding human activity - this social context is explicable at the level of the community.

To date, the problem has been defined and the process of data definition is currently underway. Thus I am still involved in the hermeneutic process; the so-called hermeneutic circle is in full swing. The GIS analysis will proceed once data have been defined to a satisfactory level which will lead to the discussion of community patterning.

4. Conclusion

Several problems in the use of GIS in archaeology have been noted. These problems have been related back to the lack of a general underlying theoretical framework from which to utilise the technology. One area in particular has been highlighted as a necessary beginning for the development of such a theory. This area is that of conceptual data modelling for the explicit definition of data to be incorporated into the analysis. A hermeneutic procedure has been outlined for this definition, and an archaeological dataset has been discussed.

References

Aldenderfer, M. and H. D. G. Maschner (eds.), 1996. *Anthropology, Space, and Geographic Information Systems.* Oxford: Oxford Univer ·ty Press.

Allen, K. M. S., S.W. Green and E. B.W. Zubrow (eds.), 1990. *Interpreting Space: GIS and Archaeology.* London: Taylor and Francis.

Batini, C., S. Ceri and S. Navathe, 1992. *Conceptual Data Modelling: An Entity-Relationship Approach.* California: The Benjamin/Cummings Publishing Company.

Bayard, D.T., 1977. Phu Wiang pottery and the prehistory of Northeast Thailand. *Modern Quaternary Research in Southeast Asia.* 3:57-102.

Bayard, D.T., 1992. Models, scenarios, variables and suppositions: approaches to the rise of social complexity in Mainland Southeast Asia, 700 BC - AD 500. In Glover, I. C., P. Suchitta and J. Villiers (eds.), pp.13-38.

Bender, B. (ed.), 1993. *Landscape: Politics and Perspectives.* Oxford: Berg.

Bietti, A., I. Johnson and A. Voorrips (eds.), 1996. *XIII International Congress of Prehistoric and Protohistoric Sciences, 8/14 September 1996. Colloquim II: The Present State of GIS Applications and Analogous Systems in Prehistoric Archaeology.* Forli, Italia.

Bishop, P., D. Hein, M. Barbetti and T. Sutthinet, 1994. Twelve centuries of occupation of a river-bank setting: old Sisatchanalai, northern Thailand. *Antiquity.* 68:745-757.

Brandt, R., B. J. Groenewoudt and K. L. Kvamme, 1992. An experiment in archaeological site location: modeling in the Netherlands using GIS techniques. *World Archaeology.* 24(2):268-282.

Burrough, P.A. and A. U. Frank, 1995. Concepts and paradigms in spatial information: are current geographical information systems truely generic? *International Journal of Geographical Information Systems.* 9(2):101-116.

Demaine, H., 1978. Magic and management: methods of ensuring water supplies for agriculture in South East Asia. In Stott, P. (ed.), pp.49-67.

Fernández, R. N. and M. Rusinkiewicz, 1993. A conceptual design of a soil database for a geographical information system. *International Journal of Geographical Information Systems.* 7(6):525-539.

Firns, P. G., 1994. *An Extended Entity Relationship Model Ap-*

plicable to the Design of Spatially Reference Databases. Dunedin, University of Otago. PhD Dissertation.

Fisher, P., 1997. Editorial. *International Journal of Geographical Information Science.* 11(1):1-3.

Flannery, K. V. (ed.), 1976. *The Early Mesoamerican Village.* New York: Academic Press.

Gadamer, H.-G., 1975[1960]. *Truth and Method.* London: Sheed and Ward.

Gaffney, V., Z. Stancic and H. Watson, 1995. The impact of GIS on archaeology: a personal perspective. In Lock, G. and Z. Stancic (eds.), pp.211-230.

Gillings, M., 1996. Not drowning but waving? - Re-humanising GIS, the Tisza flood plain revisited. In Bietti, A., I. Johnson and A. Voorrips (eds.). pp. 69-84.

Glover, I., Pornchai Suchitta and J. Villiers (eds.), 1992. *Early Metallurgy, Trade and Urban Centres in Thailand and Southeast Asia.* Bangkok: White Lotus.

Goodchild, M. F., 1992. Geographical information science. *International Journal of Geographical Information Systems.* 6(1):31-45.

Gould, M., 1994. GIS design: a hermeneutic view. *Photogrammetric Engineering and Remote Sensing.* 60(9):1105-1115.

Harris, T. M. and G. R. Lock, 1990. The diffusion of a new technology: a perspective on the adoption of geographic information systems within UK archaeology. In Allen, K. M. S., S. W. Green and E. B. W. Zubrow (eds.), pp.33-53.

Harris, T. M. and G. R. Lock, 1995. Toward an evaluation of GIS in European archaeology: the past, present and future of theory and application. In Lock, G. and Z. Stancic (eds.), pp.349-366.

Higham, C. F. W., 1977. The prehistory of the southern Khorat Plateau, with particular reference to Roi Et province. *Modern Quaternary Research in Southeast Asia.* 3:103-142.

Higham, C. F. W., 1996a. *The Bronze Age of Southeast Asia.* Cambridge: Cambridge University Press.

Higham, C. F. W., 1996b. Archaeology and linguistics in Southeast Asia: implications for the Austric hypothesis. *Bulletin of the Indo-Pacific Prehistory Association.* 14:110-118.

Higham, C. F. W., Rachanie Bannanurag, G. Mason and N. Tayles, 1992. Human biology, environment and ritual at Khok Phanom Di. *World Archaeology.* 24:35-54.

Higham, C. F. W., Amphan Kijngam and B. Manly, 1982. Site location and site hierarchy in Prehistoric Thailand. *Proceedings of the Prehistoric Society.* 48:1-27.

Ho, C.-M., 1992. An analysis of settlement patterns in the Lopburi area. In Glover, I., P. Suchitta and J. Villiers (eds.). pp.39-46.

Hodder, I., 1982. *Symbols in Action: Ethnoarchaeological Studies of Material Culture.* Cambridge: Cambridge University Press.

Hunt, E. D., 1992. Upgrading site-catchment analyses with the use of GIS: investigating the settlement patterns of horticulturalists. *World Archaeology.* 24(2):283-309.

Kvamme, K. L., 1989. Geographic information systems in regional archaeological research and data management. *Archaeological Method and Theory.* 1:139-203.

Kvamme, K. L., 1995. A view from across the water: the North American experience in archaeological GIS. In Lock, G. and Z. Stancic (eds.), pp.1-14.

Limp, W. F., 1996. Developing methodologies in the analysis of spatially referenced data and their impacts on archaeological method and theory. In Bietti, A., I. Johnson and A. Voorrips (eds.). pp.115-126.

Lock, G. R. and Z. Stancic (eds.), 1995. *Archaeology and GIS: A European Perspective.* London: Routledge.

Maguire, D. J. and J. Dangermond, 1991. The functionality of GIS. In Maguire, D. J., M. Goodchild and D. Rhind (eds.), *Geographic Information Systems: Principles and Ap-*

plications. pp.319-335. New York: Longman Scientific and Technical. Volume 1.

Mark, D. M. and A. U. Frank, 1996. Experential and formal models of geographic space. *Environment and Planning B: Planning and Design*. 23:3-24.

Molenaar, M., 1989. Towards a geographical information theory. *ITC Journal*. 1:5-11.

Moore, E., 1988a. *Moated Sites in Northeast Thailand*. Oxford: British Archaeological Reports International Series 400.

Moore, E., 1988b. La maîtrise de l'eau en nord-est Thaïlande: le changement pendant les siècles. *First Franco-Thai Symposium*, Silpakorn University.

Moore, E., 1989. Water management in early Cambodia: evidence from aerial photography. *The Geographical Journal*. 155(2):204-214.

Moore, E., 1992. Water enclosed sites: links between Ban Takhong, Northeast Thailand and Cambodia. In Rigg, J. (ed.), pp.26-46.

Mudar, K. M., 1995. Evidence for prehistoric dryland farming in Mainland Southeast Asia: results of regional survey in Lopburi Province, Thailand. *Asian Perspectives*. 34(2):157-194.

Ng, R., 1978. Man and land in Northeast Thailand. In Stott, P. (ed.). pp.34-48.

Nitta, E., 1992. Ancient industries, ecosystem and environment: special reference to the Northeast of Thailand. *Historical Science Reports, Kagoshima University*. 39:61-80.

Pigott, V., Surapol Natapintu and Udom Theetiparivatra, 1992. Research in the development of prehistoric metal use in Northeast Thailand. In Glover, I., P. Suchitta and J. Villiers (eds.), pp.47-62.

Pornchai Suchitta, 1992. Early iron-smelting in Thailand and it implications. In Glover, I., P. Suchitta and J. Villiers (eds.), pp.115-122.

Ricoeur, P., 1981. *Hermeneutics and the Human Sciences*. Cambridge: Cambridge University Press. Translator: J. Thompson.

Rigg, J. (ed.), 1992a. *The Gift of Water: Water Management, Cosmology and the State in South East Asia*. London: School of Oriental and African Studies.

Rigg, J., 1992b. The gift of water. In Rigg, J. (ed.), pp.1-6.

Rodaway, 1994. *Senuous Geographies*. London: Routledge.

Stott, P.A. (ed.), 1978a. *Nature and Man in South East Asia*. London: School of Oriental and African Studies.

Stott, P. A., 1978b. Nous avons mangé la forêt: environmental perception and conservation in mainland South East Asia. In Stott, P.A. (ed.), pp.7-22.

Tambiah, S. J., 1970. *Buddhism and the Spirit Cults in North-East Thailand*. Cambridge: Cambridge University Press.

Taylor, P. J. and M. Overton, 1991. Further thoughts on geography and GIS. *Environment and Planning A*. 23:1087-1094.

Thomas, J., 1993. The hermeneutics of megalithic space. In Tilley, C. (ed.), *Interpretative Archaeology*. pp.73-97. Oxford: Berg.

Tilley, C., 1994. *A Phenomenology of Landscape*. Oxford: Berg.

Trigger, B. G., 1978. *Time and Traditions: Essays in Archaeological Interpretations*. Edinburgh: Edinburgh University Press.

van Liere, W., 1980. Traditional water management in the lower Mekong Basin. *World Archaeology*. 11(3):265-280.

Voorrips, A., 1996. Archaeological theory and GIS, any relations? In Bietti, A., I. Johnson and A. Voorrips (eds.), pp.209-214.

Warnke, G., 1988. *Gadamer: Hermeneutics, Tradition and Reason*. Stanford: Stanford University Press.

Welch, D. J., 1985. *Adaptation to Environmental Unpredictability: Intensive Agriculture and Regional Exchange at Late Prehistoric Centres in the Phimai Region in Thailand*. University of Hawaii. Ph.D thesis.

Welch, D. J. and J. R. McNeill, 1991. Settlement, agriculture and population changes in the Phimai region, Thailand. *Bulletin of the Indo-Pacific Prehistory Association.* 11:210-28.

Wheatley, D., 1993. Going over old ground: GIS, archaeological theory and the act of perception. In Andresen, J., T. Madsen and I. Scollar (eds.), *Computing the Past: Computer Applications and Quantitative Methods in Archaeology.* pp.133-138. Aarhus: Aarhus University Press.

Wheatley, D., 1995. Cumulative viewshed analysis: a GIS-based method for investigating intervisibility, and its archaeological application. In Lock, G. and Z. Stancic (eds.). pp.133-138.

Wilen, R. N., 1987. Excavation and site survey in the Huai Sai Khao Basin, Northeast Thailand. *Bulletin of the Indo-Pacific Prehistory Association.* 7:94-117.

A GIS Based Walkway Management System

Mary Avery,
Ryan Clements
Centre for Computing and Biometrics
Lincoln University,
Christchurch,
New Zealand
averym@lincoln.ac.nz
clementr@lincoln.ac.nz

Glen Harrison,
Hughey
artment of Resource Management
Lincoln University
Christchurch, New Zealand
harrisog@lincoln.ac.nz
hugheyk@lincoln.ac.nz

Marty Thompson
Conservation Officer
Department of Conservation
Mount Thomas
Canterbury Conservancy
Private Bag
Rangiora

Presented at the second annual conference of GeoComputation '97 & SIRC '97, University of Otago, New Zealand, 26-29 August 1997

Abstract

The Walkway Management System (WMS) uses geographic information system (GIS) software to calculate an estimate for the level of maintenance required for walkway segments. It then assists the user in prioritising the maintenance on segments of the walkway that require repair. The development of the WMS is a cooperative effort between a team of researchers at Lincoln University and Department of Conservation (DoC) staff. DoC staff provided guidance and data, and the Lincoln University research team has implemented the system in ArcInfo software. This paper provides an analysis of the walkway maintenance problem and an overview of a GIS application developed for use as an applied tool for resource management.

1 Background

Outdoor recreation is a major pastime of New Zealanders and visiting international tourists. In recent years, there has been a dramatic increase in demand for wilderness[1] experiences. This demand has put tremendous pressure on the country's walking tracks (Kearsley & Gray, 1993). With the changes in patterns of visitor numbers, use and expectations, it is vital that managers plan for the future to provide appropriate services and facilities, without endangering the resources that the visitors have come to experience (Marshall, 1994).

An estimated 2.4 million visits were made to DoC offices in 1994/95. Current international visitor numbers are over one million each year, and the Tourism Board expects numbers to increase to two million by the year 2000 and three million by the year 2004. About half of these people visit areas managed by DoC (DoC, 1996a).

In April 1987, administrative changes led to the creation of the Department of Conservation. DoC assumed management of New Zealand's national parks, forest parks and other protected areas, including the numerous walkways from the Department of Lands and Survey and New Zealand Forest Service.

In September 1994, DoC published a Visitor Strategy Discussion Document (DoC, 1994). It states the Department's objectives as being:

> "(a) to protect New Zealand's natural and historic heritage
>
> (b) to provide opportunities for people to appreciate, use and enjoy the lands and waters it manages - but with care and respect
>
> (c) to act as a voice for conservation in the community and in government."

[1] We use wilderness as a relative term depending on the user's perspective. A user may consider the wilderness to be a short walk on a wooded trail near an urban area, while others may consider the wilderness to be a back country trail.

This document was written as the first step in the process of addressing the issues of management and planning for the resources under DoC's care in relation to the changes taking place in visitor flow and needs.

In October 1996, the Greenprint documents outlined DoC's policies to the incoming government (DoC, 1996a and b). *The* Visitor Strategy *in this document set five goals:*

"(a) Protection
(b) Fostering visits
(c) Managing tourism concessions on protected lands
(d) Informing and educating visitors
(e) Visitor safety."

When the documents were written, DoC was responsible for the management of about 27 per cent of the country's land area, with about 8600 kilometres of walking tracks, 1200 kilometres of roads, 960 huts, 250 campsites, 40 visitor centres and thousands of roadside, waterside and road-end facilities. Visitor structures managed by DoC include boardwalks, boat ramps, jetties, pedestrian and vehicle bridges, retaining walls, safety fences, guard rails, and viewing platforms. There are between 15-20,000 structures at 4500 sites.

The Department recognises the value of GIS in the management of these land, facilities and walkways. McEwen (1990) discussed the ways GIS could be used to assist DoC with its land and facility management problems.

DoC classifies walkways into four categories; path, walk*ing track, tramping track and route. The level of visitor use* for each walkway segment is an important consideration *in determining the upkeep of the walkway. The greater* the walkway's use, the more investment usually goes into its upkeep. Another consideration is the walkway category. Due to user needs and perception, a path requires more maintenance than a route. A path is used predominantly by families, less experienced walkers and the disabled. These users require a higher standard of walkway and facilities, and as there are more of these users there is a need for more facilities to be provided. Whereas, a route is generally used by well equipped and experienced trampers who are interested in the rough and rugged wilderness, and do not require carefully maintained walkways and facilities.

Walkway maintenance is one of the major problems that DoC has. McQueen (1991) has outlined some of the environmental impacts of visitor use on walkways. In addition, Simmons and Cressford (1989), Stewart (1985), and Young (1985) have researched the effects of the environment on walkways. Some general conclusions drawn from this research are discussed below. These conclusions are supported internationally (Department of Parks Wildlife and Heritage[2], 1994), and by the casual observation of locals and frequent walkers (Grzelewski, 1995).

One of the major areas of concern for DoC is the environmental impact of the increased visitor use on walkways. Frequency of visitor use is often one of the major causes of walkway deterioration. The higher the number of users, the greater the impact of trampling (although on gravel surfaces high user numbers compacts the substrate, lessening the need for maintenance). Other problems, such as walkway widening, occur where the walkway is congested and walkers overtake each other or where the walkway shows signs of deterioration, in which case the users will walk on the more stable edges of the walkway. Unplanned walkway formation occurs when users go off the designated walkway creating a new walkway through formerly untracked areas. This can lead to locally severe environmental impacts, as well as lowering the recreational and wilderness value of the area.

Other factors such as slope, aspect, soil type, rainfall, walkway surface and vegetation influence the rate of walkway deterioration. Walkways on steeper slopes tend to have water flowing off the slope over the walkway causing erosion. Walkways on flat surfaces may have drainage problems. High intensity rainfall has a more detrimental impact on walkways than low intensity rainfall. Organic soils are more susceptible to damage than gravel soils. The north, west, and northwest aspects receive more impact from

2 The Tasmanian Parks & Wildlife Service has developed a management strategy document. The Lincoln research team has been in contact with the Tasmanian Parks & Wildlife Service and we will be sharing ideas and results with them.

wind during the year than the other directions. All these factors and others need to be considered in the management of walkways.

The Mount Thomas and the Oxford Forests in North Canterbury were selected for use in the WMS prototype development. The Mount Thomas Forest, located 60 kilometres northwest of Christchurch, covers an area of 10,800 hectares. It has six walkways of varying length, a picnic/camping area, permanent fire places, toilets and running water. The Oxford Forest, located approximately 56 kilometres from Christchurch, covers an area of 11,350 hectares. It has four walking tracks and four tramping routes of varying length (DoC, 1991). These two sites were chosen for their proximity to Christchurch, the number of walkways and facilities associated with the area, the available data, and the availability of local knowledge to assist in the development of the prototype.

2 System Development

2.1 Problem Definition

DoC is in the unenviable position of having to balance the need to protect the environment and resources for which it is responsible with the desires of the recreational visitors who wish to use those very resources. In making management and planning decisions, DoC must keep these two apparently opposing needs in mind.

Due to the limited funding that DoC receives and the large number of facilities, services and lands it has to manage and maintain, there is a need for DoC to efficiently allocate its limited financial resources. Currently, DoC uses a

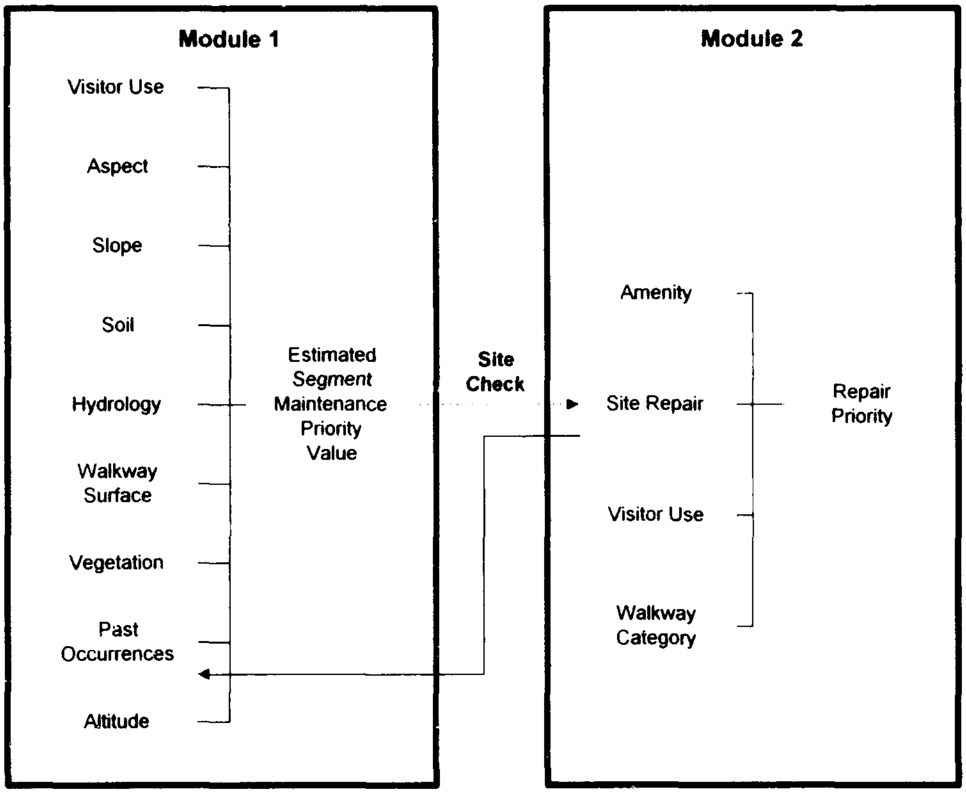


Figure 1 - WMS Prototype Maintenance Model Structure Diagram

combination of manual and automated techniques to evaluate the need for walkway maintenance and repair. No one system has the information required to make a standard and efficient evaluation of walkway maintenance priorities.

2.2 Problem Solution

The WMS prototype was implemented primarily in ArcInfo GIS software. GIS provided the functionality to analyse the spatially coexistent factors that impact upon walkways. In the early stages of the conceptual design, the research team recognised that modeling the physical factors could only provide a range of probabilities for maintenance on walkway segments. There needed to be a knowledgeable observer to then evaluate these segments for actual maintenance needs. The actual maintenance requirement could then be input into the system and a prioritised maintenance ranking would be generated based on walkway characteristics and use. Figure 1 illustrates the conceptual solution consisting of two principal modules.

Module One calculates an Estimated Segment Maintenance Priority Value for each walkway segment based on its level of visitor use, aspect, slope, soil type, hydrology, vegetation, track surface, altitude, and past maintenance characteristics. The higher this value, the greater the likelihood that this segment will require maintenance. This result gives the user a set of rank order track segment locations where maintenance problems would most likely exist. These results provide an indication of the resources needed to inspect the walkway network for required maintenance and potential maintenance needs.

Module Two maintains information on the required maintenance or repair. Needed repairs are input into the Site Repair component of the Repair Priority module (Module Two). This is done from a pick list of different categories of maintenance required. The amenity value (Archaeological Sites, Species Index, Areas of Natural Significance, Geological Preservation Sites), site repair value, walkway category and level of visitor use values are combined and sorted to provide the user with a segment repair priority listing. Armed with this information, the user can then determine a walkway maintenance schedule.

3 Prototype Implementation

Most of the digital geographic data required for the prototype was held by the DoC Canterbury Conservancy Christchurch office in Terrasoft GIS format. Data such as contours, walking tracks, streams and soil and vegetation polygons were converted to ARC/INFO format by DoC staff. The Lincoln University research team then manipulated the base data layers to include only the information relevant to walkway maintenance. These layers are the maintenance factors in the WMS prototype.

The item's (database fields) Factor Class, Factor Value and Factor Weighting were added for each maintenance factor and populated with data. These values were discussed with DoC experts and adjusted based on their input.

ArcView was used for display and query purposes. This software was chosen because of its relative simplicity and availability at DoC conservancies. The ability of DoC users to query attribute information and produce maps of the walkway network was considered to be important.

Both modules required graphical display of results. Walkway segments were colour coded to indicate priority. Maps can be simply produced to show the location and rank of all track segments or to highlight only those which have been designated within the highest priority range.

3.1 Module One

Slope and aspect polygons were derived from 20 metre interval contour data. A 50 metre resolution lattice was created from a TIN of the study area which provided appropriately generalised slope and aspect information. Walkway visitor numbers were obtained from DoC field records and linked to the walkways by walkway site number. Walkway surface attributes were manually attached to walkway segments. A hydrology coverage was created by buffering streams to a distance of five metres.

Maintenance factors were combined using line-in-polygon overlay to produce a segmented walkway coverage. Walkway segments varied in length from tens of centimetres to tens of metres depending on the variation in visitor use,

aspect, slope, soils, hydrology, walkway surface, vegetation, and altitude.

A model that sums the maintenance factors was developed using the following equation (factor values and factor weighting variables are defined in Table 1).

Estimated Segment Maintenance Priority Value = [(Fvu * Wvua) + (Fa * Wa) + (Fsl * Wsl) + (Fs * Ws) + (Fh * Wh) + (Fws * Wws) + (Fv * Wv) + (Fal * Wal) + (Fpo * Wpo)]

The result is a numerical maintenance priority value for every walkway segment. These priority values are sorted and grouped into classes for display.

3.2 Module Two

Module Two operates on the same segmented walkway coverage as Module One (only necessary attributes were retained). Actual repair event data are added by selecting the location graphically and inputting a site repair value and a description of the repair required using an input form.

Input of repair events is obtained through the use of a pick list of different categories, such as trees over the walkway, landslide, and walkway wash-out. Each of these categories has a different value based on the degree of walkway blockage that they cause. Amenity values are given to each walkway segment leading to a specific amenity.

In addition to actual repair events, statutory site inspection requirements are incorporated. These are assigned site repair values such that they would rank the highest. Those walkway segments that have site inspection requirements assigned to them are displayed in a separate category.

A model was developed that sums this repair data with walkway usage, walkway category and amenity value to calculate a repair priority value using the following equation (factor values and factor weighting variables are defined in Table 2).

Site Repair Priority Value = [(Fam * Wam) + (Fvu * Wvub) + (Fwc * Wwc) + (Fsr * Wsr)]

The results of this equation are displayed on a colour coded map to show the ranking of the walkway segments by repair priority.

4 DoC Feedback and Field Test

The results from an initial test run were used by the research team to review the system with DoC staff at Mount Thomas. The structure of the system, the factors that should be used in each module, the factor values, and weight values were all reviewed. Whilst the initial results were deemed to be reasonably accurate, a number of factor values and weightings were revised, along with the factors and their categories. A similar meeting was held with DoC management staff at the Canterbury Conservancy office in Christchurch where additional suggestions were made. Both field and management staff could see the potential value of the system for their respective long term planning and day to day implementation of maintenance. Interest was expressed, without formal commitment, to see full implementation of the system.

The results of the WMS prototype were field tested on the Mount Thomas Forest tracks. Researchers found that maintenance priority values should have been higher where introduced vegetation species occurred and in areas of southwest aspect. Introduced plant pest species result in consistent problems of encroachment on the walkway. The snow on the southwest aspect of the hills, which had not been taken into account has caused considerable damage to trees along walkways in years past.

The changes from the discussions and field test were noted and incorporated into the system. The results generated by the revised WMS prototype were more realistic and useful.

5 Assumptions and Limitations

Visitor numbers are taken by DoC as one way traffic. This has major implications for the amount of deterioration on a walkway due to visitor use. For instance, if the walkway is a single return route, the visitor would be counted once, even though the trail would have been traverse twice by

Table 1 - Evaluation Tables used in the Prototype Maintenance Model: Module One

Maintenance Factor	Factor Class	Factor Value	Factor Weighting	Description
Visitor Use (Fvu * Wvu)	0 - 499	1	10	* little impact
	500 - 999	2		* walkway deterioration
	1000 - 1999	3		* moderate walkway Deterioration
	2000 - 2999	4		
	3000 - 4999	5		* almost total soil removal
	5000 - 9999	6		
	10000 - 19999	7		* severe walkway Deterioration
	>=20000	8		
Aspect (Fa * Wa)	North	2	5	* Number represents the level of impact from rain, wind and Snow
	North - East	1		
	East	1		
	South - East	0		
	South	1		
	South - West	2		
	West	3		
	North - West	3		
Slope (degrees) (Fsl * Wsl)	S <= 1	5	3	* flat to gentle
	1 > S <= 2	5		
	2 > S <= 3	5		
	3 > S <= 5	5		
	5 > S <= 10	1		* gentle to moderate
	10 > S <= 20	2		* moderately steep
	20 > S <= 35	3		* steep
	35 > S <= 55	4		* very steep
	55 > S <= 90	5		* precipitous
Soil (Fs * Ws)	Gravel soils	1	2	* low impact
	Associated Yellow-brown shallow & stony soils	1		
	Yellow-grey earths	2		* moderate impact
	Yellow-grey to Yellow-brown earths intergrade	2		
	Lowland Yellow-brown earths	2		
	Upland & high country Yellow-brown earths	3		* high impact
	Recent soils	3		
	Organic	4		* severe impact
Hydrology (Fh * Wh)	> 5 metres	1	1	Potential for washing out and erosion from stream overflow and flooding
	< 5 metres	2		
Walkway Surface (Fws * Wws)	Rock	0	3	* little impact
	Top Course	1		* compacts down
	Natural	3		* top soil and vegetation easily Impacted
Vegetation (Fv * Wv)	Alpine Tussockland	1	2	* high durability to trampling
	Grassland	2		
	Beech Forest	3		* maintenance required
	Broadleaf Forest	4		
	Introduced - all types	5		* high maintenance required
Past Occurrences (Fpo * Wpo)	0	0	20	Number of past occurrences increases the potential of Occurrences
	1-2	1		
	3 - 4	2		
	> 4	3		
Altitude (Fal * Wal)	>= 750 metres	0	7	The higher the altitude the less maintenance required
	< 750 metres	1		

Table 2 - Evaluation Tables used in the Prototype Maintenance Model: Module Two

Repair Factor	Factor Class	Factor Value	Factor Weighting	Description
Amenity (Fam * Wam)	Only w[illegible] or mo[illegible]	4	2	All sites of significancesuch as archaeological sites, areas of natural significance, geological preservation sites and species index
	Only way to 1	3		
	Shared way to 2	2		
	Shared way to 1	1		
	None	0		
Visitor Use (Fvu4 * Wvu4)	0 - 499	1	4	The more users the greater the priority for maintenance
	500 - 999	2		
	1000 - 1999	3		
	2000 - 2999	4		
	3000 - 4999	5		
	5000 - 9999	6		
	10000 - 19999	7		
	>=20000	8		
Walkway Category (Fwc * Wwc)	Route	1	1	Expectations and level of experience differs from Path users to Route users therefore needs for quality of walkway and facilities differ
	Tramping Track	2		
	Walking Track	3		
	Path	4		
Site Repair (Fsr * Wsr)	Fallen tree: Minor	2	5	Tree has fallen on walkway. Walkway still useable.
	: Major	10		Tree has fallen on walkway. Walkway impassable.
	Landslip : Minor	2		Small slip. Walkway still useable with little or no danger.
	: Major	10		Major slip. Walkway closed due to danger to users.
	Washed out: bridge	10		Walkway impassable.
	: walkway	10		Walkway impassable.
	Damaged: stairs	1		Stair broken or damaged. Walkway still useable.
	: bridge	2		Bridge broken or damaged. Walkway still useable.
	: boardwalk	1		Boardwalk broken or damaged. Walkway still useable.
	: platform	3		Platform broken or damaged. Walkway still useable.
	Tree roots	2		Tree roots damaging walkway. Walkway still useable.
	Flooding	4		Walkway or structure flooded. Walkway still useable.
	Site inspection	200		Mandatory site inspection.

the person walking up and back. This highlights the need for more precise visitor monitoring to fully gauge the actual number of people walking on each segment.

Some of the data in the current tables have been developed from studies of other areas, localised information sources and input from local DoC staff. More research needs to be done to confirm the relationship between the physical factors and track maintenance, so that the results obtained for the Estimated Segment Maintenance Priority Value more closely reflect reality. User feedback will also be necessary from actual operational experience to adjust the factor values and factor weights to ensure the great-

est model accuracy.

Generalisations were made for some of the physical factors that may not be valid for an expanded area of analysis. For instance, due to the relatively small size of the current study area, it is assumed precipitation is constant. The impact of precipitation is taken into consideration by using the walkway surface and category, slope, soil, hydrology, and vegetation factor values. Precipitation variation will need to be used if the WMS is applied to a wider area.

Data input for actual repair events is associated with walkway segments, rather than point locations. This may result in accuracy problems for longer segments.

6 Implications and Further Development

DoC staff can use the prototype to more efficiently apportion their resources for maintenance and repair activities on walkways. The system can be used for both long-range planning or short-range evaluation of priorities.

The WMS prototype uses the Mount Thomas and Oxford Forests as a test case. After the system is refined, there is potential to expand it to cover more areas managed by DoC (e.g. conservancy or nationwide).

The system could provide an estimate of the cost for repairs and maintenance based on a standard set of costs for different categories of work. This would then enable DoC staff to quickly determine not only priority, but total cost. A report on specific maintenance that is needed could also be sent from the WMS to project planning software for efficient scheduling of these tasks.

If new or altered walkway construction is planned, an extension of the WMS software could be used to determine estimated maintenance requirements based on the physical features of the land and the estimated visitor use. DoC could use this data to manage the tradeoffs between maintenance costs and provision of access to walkways.

For the long term, WMS could be incorporated into a broad based GIS Walkway Management System (WMS) that could include an interactive visitor interface. This visitor interface could provide information on walkway category, level of use, current walkway conditions, distances and average walking times for the walkway, equipment required, recommended experience level, points of interest along the walkway segments, and map printouts.

7 Conclusion

The Walkway Management System prototype is a first attempt to model the complex physical and human factors that result in maintenance needs on the different categories of walkways. GIS has already been used to record maintenance needs for transportation infrastructure, but this research extends GIS capabilities beyond a record's management function to provide an analytical and management tool that can be used for short term and long term decisions for walkway management, maintenance and viability.

Acknowledgments

The authors wish to acknowledge the contribution of DoC staff in providing data, advice and time for the development and refinement of the WMS prototype. In particular, we would like to thank Norm Thornley, Dean Strachan and Jon Bos at the DoC Canterbury Conservancy office in Christchurch. We would also like to acknowledge the Centre of Mountain Studies via the Struthers Legacy at Lincoln University for funding the development of the WMS prototype.

References

Department of Conservation. (1991) *North Canterbury's Foothill Forests: A Recreational Guide.* Department of Conservation.

Department of Conservation. (1994) *Visitor Strategy Discussion Document.* Department of Conservation.

Department of Conservation. (1996a) *Greenprint Conservation in New Zealand - a strategic overview, Volume one.* Department of Conservation.

Department of Conservation. (1996b) *Greenprint Conservation in New Zealand - a guide to the department,*

Volume two. Department of Conservation.

Department of Parks, Wildlife & Heritage (Tasmania). (1994) *Walking track management strategy for the Tasmanian Wilderness World Heritage Area.* Tasmania. Department of Parks, Wildlife & Heritage.

Grzelewski, D. (1995) Routeburn. *New Zealand Geographic,* 25, 18-32.

Kearsley, G.W. and Gray, G. (1993) International visitor flows and infrastructure needs: a New Zealand example. *New Zealand Geographical Society Conference 1993.*

Marshall, D. (1994) In *Visitor Strategy Discussion Document.* Department of Conservation 1994.

McEwen, M. (1990) *GIS and the DoC.* Department of Conservation.

McQueen, D. (1991) *Environmental impact of recreational use on DoC estate. Part 1. Guidelines on track location, management and repair.* DSIR Land Resources Contract Report No. 91/54 (part 1).

Simmons, D. G. and Cressford, G. R. (1989) *The St James Walkway Study.* Occasional Paper 1. Dept Parks, Rec, Tourism. Lincoln University, New Zealand.

Stewart, D. P. C. (1985) *Effects of trampling on the soils of the St James Walkway, Lewis Pass.* Unpub. Dissertation B.Ag.Sc (Hons), Lincoln University, New Zealand.

Young, J. R. (1985) *Erosion control of New Zealand walking tracks.* Hons thesis, Lincoln College, Univ. of Canterbury, New Zealand.

New Zealand Land Tenure Beyond 2000: Full Integration and Automation of the New Zealand Survey and Title Systems

Don Grant, Tadeusz Dawidowski & Ralph Winmill
Land Information New Zealand
PO Box 5501
Wellington
New Zealand

Presented at the second annual conference of GeoComputation '97 & SIRC '97, University of Otago, New Zealand, 26-29 August 1997

Abstract

Land Information New Zealand has been charged with the development of a strategy for the integrated automation of the survey and title systems. This is a new programme with new management structure and new objectives. The preliminary phase to determine user requirements has been granted funding approval by Cabinet. Following completion of this phase, approval for the rest of the programme will be sought.

This paper explores the principles, impacts and opportunities of this new integrated system from a survey perspective. The automation strategy will involve a redesign of systems and processes to allow the full benefits of automation to be realised.

A fundamental principle of this concept is that the survey and title transactions will merge into a single digital land transaction. This will enable surveyors and solicitors to develop new relationships for creating and submitting transactions in land.

The impact of an integrated land tenure system on the existing survey and title systems is one of complete process automation with the implied digital conversion of "physical records". This digital conversion would not simply be a change of format from paper to static digital records such as scanned plans (although it may include this for historical records). It would also involve creation of live and intelligent digital records that play an active role in automated processes.

This automation strategy will not only retain the principles of the survey and title systems, but will extend them and completely alter the way in which they operate. It will also enable Land Information New Zealand to meet its vision of providing world class land and seabed information services.

Introduction

Background

Prior to the restructuring of the former Department of Survey and Land Information (DoSLI), Survey System management embarked on a programme of change for the current survey system. The primary drivers were to:

- reduce costs;
- improve efficiencies;
- meet changing requirements of the National Spatial Reference System; and to
- ensure that the survey system could take full advantage of developing technology capabilities (which in turn dictate new user requirements).

Analysis confirmed that the current survey system is reaching its limit of cost-effective improvement and would not be able to meet the envisaged needs of the users of the 21st century and beyond.

The restructuring of DoSLI, and the subsequent creation of Land Information New Zealand resulted in this programme being reassessed in terms of the new Department's vision and business drivers

The New Department - Land Information New Zealand

The restructuring and refocusing of the former DoSLI as Land information New Zealand was designed to ensure the effective and efficient delivery of public good land - related services in order to maintain and accelerate New Zealand's economic growth. The Chief Executive and staff identified the following business drivers for the new Department (Land Information NZ, 1996a):

- focus on core business functions of maintenance and provision of core data, processes and information
- improve Department efficiency and effectiveness
- fully integrate the former Land Titles Office and DoSLI functions
- contract non-core functions
- provide a platform for 3rd party services

Government Outcomes - Survey and Titles Responsibilities

The principle functions which must be undertaken by the Survey and Tiles systems to meet the Government's requirements are set out in Figure 1.

This illustrates that, in addition to specific Crown Related services, Land Information New Zealand's principle function is the management of core land information.

SURVEY AND TITLE AUTOMATION

In order to meet the Department's business drivers and Government outcomes concerning the management of land information, a Survey and Title Automation Strategy project was commissioned. (LINZ, 1996a)

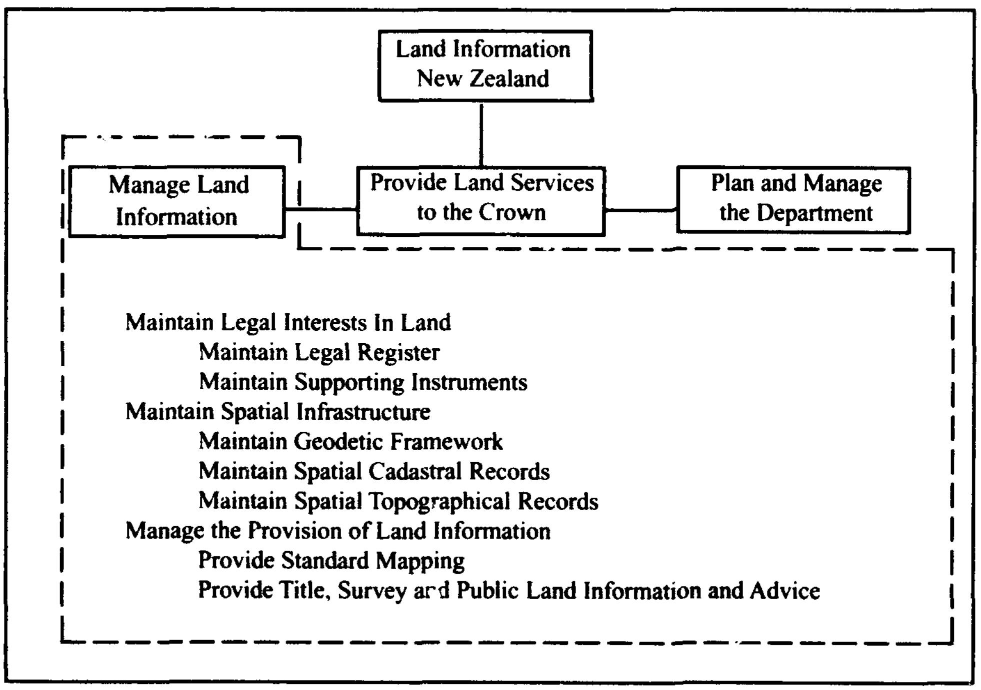


Figure 1 Land Information NZ Business Functions

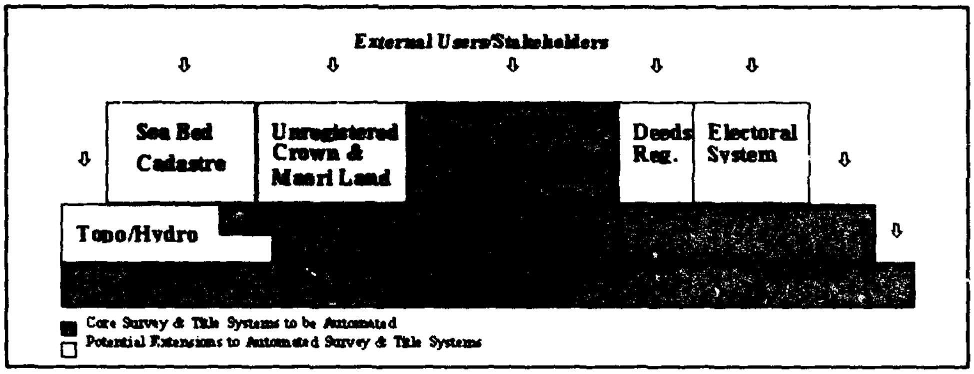


Figure 2 Core components of the proposed Automated Survey & Title Systems and users / stakeholders interests

Automation Systems Strategic Vision

The vision is of a fully digital information systems environment within the Department which is closely integrated with external users of land information. The vision recognises that information is a key strategic resource for the Department and therefore the exploitation of the capabilities of technology and the adoption of 'best practice' in information management are pivotal to the success of the new Department meeting its business objectives.

The diagram shown in Figure 2 represents a high level view of the Survey and Title Automation building blocks and the relationships to other LINZ businesses as well as users/ stakeholders.

Strategy - Phase One

This stage consists of projects which will review current legislation, define the records which will be core to the Survey & Title Automation Programme, obtain users requirements of the geodetic, cadastral survey and title systems and proceed with initial data and process analysis.

The two objectives of this stage are:

- Obtain sufficient information to present a comprehensive business case to government for approval of stage two funding.
- Identify and specify the business needs, based on user requirements, in order to provide the main input for the subsequent design and build projects.

Design Core Land Record Project

This project has been completed and the following Entity Relationship Model (also known as a Data Model) illustrates the major business entities and their relationships to one another. It is a conceptual definition of the target Survey and Title Core Land Record.

Key Points of the Core Land Record Model:

The Core Land record model:

- supports the vision of an integrated Survey and Title record through a single data model
- supports the transition period from a paper-based to digital system
- the conversion of paper records to "intelligent" records is the key enabler for process re-design
- includes the automation and redefinition of the business rules/processes which will realise the primary benefits
- provides for the survey plan and title to be seen as views of the digital data set.

Strategy - Phase Two

On funding approval the design and build projects of the Core Record Information Management System (CRiMs) and Geodetic Management System (GMS) will commence. In addition, the projects which are required for data conversion will be defined and implemented. It is envisaged, that this will include scanning, conversion, reformatting and back capture of all the required data from paper and existing digital records.

Impact from a survey perspective

The building blocks defined in Figure 2 indicate that the Survey System will have a significant and active role in the management of the Department's Core Land Record, as well as providing a spatial infrastructure for its other users. The following sections discuss the principles and, user requirements, impacts and opportunities of a New Zealand's Geodetic survey system.

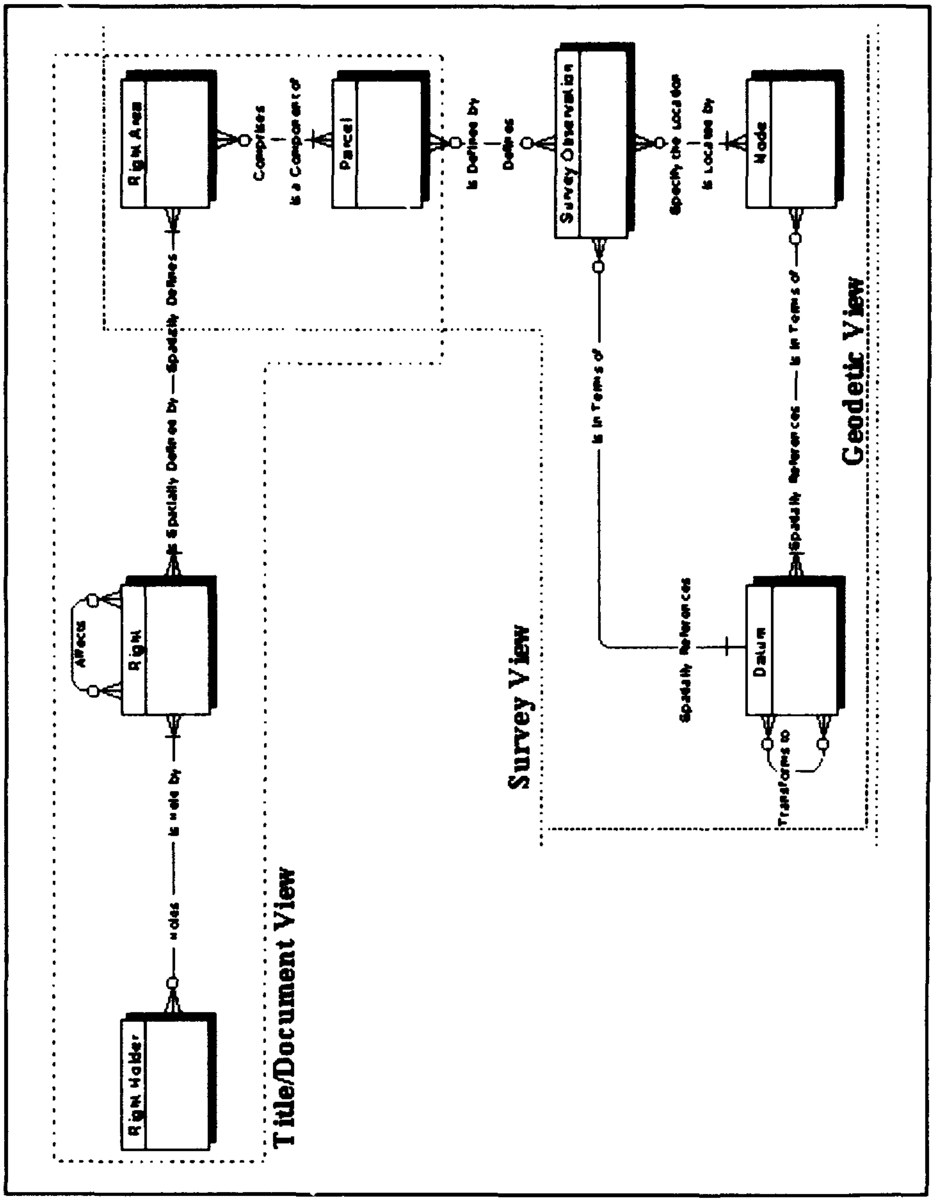


Figure 3 Entity Relationship Conceptual Data Model (Land Information NZ, 1996b)

Automated Core Record Information Management system

Fundamental Principles

With the proposed automation of Land Information NZ's Survey and Title business one of the basic principles that will not be compromised is the Department's primary role of protecting the land tenure system and the "state guarantee of title". Also any changes to the management of cadastral data should not diminish the integrity of the data that the Department holds.

Accuracy Standards

The proposed Survey Regulations (1997) have been prepared to be "output oriented" rather than "prescriptive" and process driven. However there are accuracy, and monumentation standards in the Regulations that the cadastral surveyor is required to comply with before a survey dataset is accepted for integration into the Department's authoritative spatial record. The proposed new Survey Regulations and accreditation of surveyors to undertake cadastral surveys in NZ will make surveyors more accountable for the quality of the data submitted and Land Information NZ will focus more on maintaining the integrity of the survey system. Land Information NZ will seek to be responsive to the "intent" and quality of the data lodged rather than its "legal form".

The professional surveyor's prime responsibilities will be to ensure that the survey accuracy, survey definition, and record completeness of data lodged meet the required standards and the Department will be responsible for the accuracy standards, the integration of survey datasets into its spatial record, and for the integrity of those records.

Random and routine audit procedures comprising field survey inspections and office data examination will be undertaken by Land Information NZ to verify that compliance with the accuracy standards has been achieved by the surveyor and to support the proposed accreditation system.

Survey Accurate Coordinate Cadastre

The fundamental building block for the survey component of an automated cadastral Core Record Information Management System is a coordinate network that will allow efficient electronic validation of new survey data. Crucial to this validation process is a requirement that there be a national geodetic control framework in place to underpin the integration of all cadastral survey data into a single database. The accuracy of any set of coordinates can only be as good as the coordinate system that they are derived from so in an efficient automated environment cadastral survey datasets need to have their coordinates derived from the geodetic system.

A clarification needs to be made here that a survey accurate coordinate cadastre does not give the coordinate any legal significance, or status, and the hierarchy of evidence of the physical "monument in the ground" still takes precedence over its coodinated value. The coordinate will not constitute cadastral evidence in its own right.

The coordinates provide a summary of survey data that will enable existing survey marks to be more easily found and verified. In conjunction with other survey evidence the coordinates may allow for boundary monuments to be reinstated. However, the historical survey data still remains the core evidence of establishing, and verifying, boundary location. It is not necessary, or desirable, for the role of the boundary monument to be changed by automation.

This supports the principle that in the case of a disagreement with the Core Record Information Management System (being a representation or summary of the survey data), the historical survey data, presently in the form of approved plans , or in the future - digital transactions, will remain the core evidence of boundary location.

Geodetic Management System

User Requirements

The following preliminary conclusions have been drawn from discussions with users of the Land Information NZ geodetic System:

- Many users require a GPS compatible geodetic datum.
- Spatial accuracy requirements are often higher than

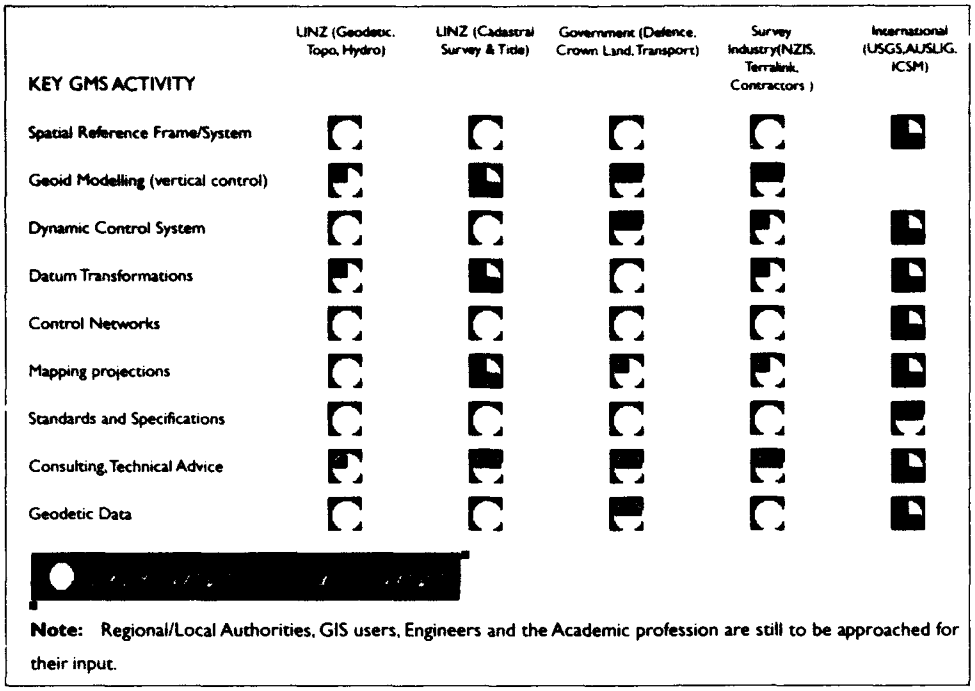


Figure 4 User Requirements Summary

currently provided by New Zealand Geodetic Datum 1949.

- Although horizontal positioning is the main requirement of the geodetic system, there is a continuing requirement for orthometric heights and an increasing need for three dimensional positioning incorporating ellipsoidal heights.
- Reduced geodetic observations will need to be held on-line to allow:

efficient validation and integration of new geodetic data

generation of up-to-date and accurate coordinates on request

maintenance and application of velocity models.

Figure 4 is a high level summary of the feedback obtained.

New Zealand Geodetic Datum 2000

Grant (1995) outlines some of the options for defining a new geodetic datum which can maintain accuracy in the presence of continuous and pervasive earth deformation.

The proposed "dynamic" datum will have the following design features:

- Dynamic modelling is necessary for continued automated processing of geodetic data and continued maintenance of coordinate system accuracy.
- As cadastral survey definition is based on boundary marks, the coordinates of these marks, and the supporting geodetic control marks, must necessarily change to reflect earth deformation.
- Maintenance of long range accuracy (made possible by GPS and dynamic modelling) will reduce the need for survey origins to be obtained locally. This, in turn, will facilitate efficient use of GPS base stations for cadastral and other surveys.
- The combination of 3D ellipsoidal coordinates and a geoid model will enable continued maintenance of the vertical control system without expensive conventional levelling.
- It is proposed that the new datum will be implemented

at a national network level (Zero, First & Second Order 2000) by 1 July 1998 and that the reference epoch for coordinates will be 1 January 2000. A national velocity model will enable data to be transformed to and from this reference epoch as required.

New Geodetic 2000 Network

Blick & Linnell (1997) outline the general features of the new geodetic network which will make New Zealand Geodetic Datum 2000 available to support the Survey & Title Automation Programme. Its features include:

- Accessible stations (generally drive-on access).
- A complete connection has been made to New Zealand Geodetic Datum 1949 1st Order marks to enable accurate transformation of historical data to the new datum.
- The horizontal & vertical networks have been integrated through observation of selected benchmarks by GPS.
- The network density for 3rd & 4th Order 2000 stations is primarily driven by cadastral survey requirements.
- The Number of geodetic "Orders" in the hierarchy may be reduced in the future as GPS cost/accuracy equation becomes less dependent on distance.

Conclusions

The Land Information NZ Survey & Title Automation Programme is currently at the stage of system analysis and definition of user requirements. The design, build and populate stages will depend on government acceptance of the business case to be presented.

The programme envisages full integration of the geodetic survey, cadastral survey and title systems. A single conceptual data model is being developed for the existing three systems and business process models will be aligned wherever practicable. This will allow Land Information NZ to realise internal cost savings in undertaking its functions and will also deliver significant savings to external users.

The efficient automation of survey data (geodetic or cadastral) will depend on provision of an accurate and accessible coordinate system. In particular, automation of cadastral survey data and processes will be increasingly reliant on an accurate geodetic infrastructure as it this system which enables the efficient association and management of digital spatial cadastral data.

Efficient automation of geodetic, cadastral survey title processes will also require intelligent data (digital data containing attributes which enable automated "business rules" to be applied). It is envisaged that intelligent records will be generated by back-capture of historical paper or digital records and, ultimately, digital lodgement of geodetic, cadastral survey and title transactions.

References

Blick, G.H., Linnell, G., (1997) The Design of a New Geodetic Network and Datum for New Zealand, Presented at the First Trans Tasman Surveyors Conference 12-18 April, 1997, Newcastle, Australia.

Dawidowski, T.A., Burgess, P., (1996) A Paradigm Shift - Surveying in a Digital Environment, Presented at 37th Australian Survey Congress, Perth, Australia.

Grant, D.B., (1995) Accommodating Change: Development of a Dynamic Geodetic Datum for New Zealand, Presented at NZIS Annual Conference, Christchurch, New Zealand, 21-23 October, 1995

Land Information NZ (1996a) Survey and Title Automation Strategy

Land Information NZ (1996b) Design Core Land Record System Abstract

Land Information NZ (1996c) Geodetic Management System Preliminary User Requirements v1.0

Land Information NZ (1996d) Core Record Information Management System Preliminary User Requirements v1.0

An Evaluation of Digital Elevation Models for Upgrading New Zealand Land Resource Inventory Slope Data.

James R.F. Barringer & Linda Lilburne
Landcare Research NZ Ltd.
P.O. Box 69, Lincoln 8152
New Zealand

Presented at the second annual conference of GeoComputation '97 & SIRC '97, University of Otago, New Zealand, 26-29 August 1997

Abstract

Slope is a key environmental parameter which influences land use and erosion hazard. Digital elevation models (DEMs) are often used to map important topographic parameters such as slope. However, the quality of such maps depends on the quality of the DEM's representation of the earth's surface. In many cases errors in this representation are neither measured nor estimated. In this paper a real-time differential GPS is used to acquire ground truth data. This ground truth is compared with DEMs generated from contours. This analysis shows that three commonly used contour-based interpolation procedures all produce good quality DEMs.

When considering the replacement of more traditional slope maps based on field mapping or air photo and contour interpretation with DEM-derived slope maps, it is important to establish that DEM-derived slope maps do represent an improvement on existing approaches. This paper compares field mapped and DEM-derived slope maps with slopes calculated from GPS elevation data. It shows that DEMs can provide both improved spatial resolution and increased accuracy in slope maps.

1. Introduction

The New Zealand Land Resource Inventory (NZLRI) has been the primary source of land resource information for New Zealand since the early 1970s. The data in the NZLRI came from field mapping. Areas of relatively homogeneous land (polygons) were defined using aerial photographic interpretation, topographic maps and field survey. For each polygon the following attributes were recorded: rock type, soil, slope, vegetation, erosion and land use capability classification. Although the NZLRI has been stored in digital form in a Geographic Information System (GIS) since 1973, the database structure has retained its original "paper-map" format, as a single geographic layer with multiple attributes. Restructuring the database to better utilise current GIS analytical capability has been hindered by the difficulty of separating key attributes from the existing single layer, and/or the cost of remapping individual attributes. Landcare Research has identified the potential for technologies such as remote sensing (Dymond, 1992b, 1995a; Wilde, 1996) and digital elevation models (Dymond, 1992a, 1994, 1995b) to be used in operational mapping or updating of the database but they have not yet been utilised widely.

Slopes derived from digital elevation models (DEM) could be used to upgrade the slope attribute currently stored in the NZLRI. However, most DEMs are interpolated from the most commonly available source of topographic data - digital contours which in turn have been generated photogrammetrically from aerial photographs. In many cases there is no quantitative assessment of DEM accuracy, and error propagation to secondary parameters such as slope and aspect is not addressed (Fryer, 1994).

In this paper we investigate the development of a raster layer of slope data to replace the classified attribute re-

corded in the original NZLRI polygons. In particular, we review measures for determining DEM accuracy, and investigate the magnitude of errors in slope as calculated from DEMs. We also analyse the relative merits of data collected using field survey methods and DEM-based slope maps.

2. Measuring Accuracy

2.1 DEM

There are many potential sources of errors in DEMs. Contours are the most common form of topographic data from which DEMs are derived. Contours are derived using photogrammetric methods. For 8x10 inch photography gathered at 1:50000 scale these methods can lead to heighting errors of ± 0.6 m for spot heights, and ± 0.7 m for contours just from random errors in the photogrammetric process (Fryer, 1994). This could lead to contour displacements of 140 m on a flood plain with 0.25° (0.5%) slope. Most mapping organizations only guarantee that contour lines are correct horizontally to within half the horizontal interval between the contour lines 90% of the time (Fryer, 1994).

To determine DEM accuracy, we need some independent knowledge of the topography to determine the difference between the digital surface and the real elevations of the same locations on the ground. This requires both a suitable sample of ground truth points, and suitable statistics from which to derive error terms (Monckton, 1994). Most commonly such ground truth points are taken from the same topographic database as the contours, in the form of local spot heights recorded at trig stations and local peaks. However, trigs and spot heights do not provide a good sample of the landscape since they over-represent peaks, under-represent low areas, and may be non-randomly distributed (ie., biased towards hilly areas). Acquisition of ground truth points should preferably be derived by independent survey, either photogrammetric (eg., Fryer, 1994), or traditional field survey (eg., Monckton, 1994). A new method of obtaining ground truth points is by using Global Positioning Systems (GPS) which estimates position (easting, northing and elevation) from satellites.

The root mean square error (*RMSE*) between DEM and ground truth elevations can be used to measure DEM accuracy :

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n} d_i^2}{n}} \qquad (1)$$

where n = number of points
d_i = $z_{ground,i} - z_{dem,i}$
$z_{ground,i}$ = ground elevation recorded at point i
$z_{dem,i}$ = DEM elevation at point i

Alternatively Li (1988) advocates the use of the standard error (*S*) and mean error ($\bar{d}$)

$$S = \sqrt{\frac{\sum_{i=1}^{n}(d_i - \bar{d})^2}{n}} \text{ where } \bar{d} = \frac{\sum_{i=1}^{n} d_i}{n} \qquad (2)$$

RMSE is the more widely used statistic but assumes a zero mean error, and therefore no systematic bias in the DEM (Li, 1988). Both Li (1988) and Monckton (1994) suggest that this assumption is not justified.

2.2 Slope

As with DEM elevation, slopes calculated from a DEM surface are subject to several sources of error. Skidmore (1989) provides an analysis of the algorithmic accuracy of six methods for calculating slope and aspect. However, algorithmic accuracy is only one source of error in calculated slope. An important issue that does not appear to have been addressed is calculating how elevation errors propagate through slope calculations (Fryer, 1994). This may be because slope maps, while easy to produce, can be difficult to reconcile with field measurements of slope (Dymond, 1994). This is because field measurements of slope are usually "integrated" over "slope length" by an observer, whereas DEM slope is generated for a fixed slope length which is related to the sampling interval (ie., the DEM resolution). Some degree of integration over slope

length is vital to avoid "noise" from micro-topographic variations of slope which could only be mapped at very large scales. However, there are no recognised standards for defining slope length. As a result, analysis of slope errors presents significant problems because the accuracy of any ground truth slope data is unknown. Ground truth data have been derived through manual interpretation of contour data (Skidmore, 1989). Such data may be useful for testing algorithmic accuracy, but seem a questionable source of ground truth. Hammer (1995), and Dymond (1994) used detailed ground survey to locate grid points for each DEM cell centre and/or manually measured slope by clinometer to gather independent ground truth data.

Calculated slope data can be compared to measured ground truth slope data in a number of ways. Dymond (1994) used a graphic interpretation with associated trend or correlation statistics. Skidmore (1989) used Kendall's tau measure of association and Spearman's rank correlation coefficient to test for a significant positive correlation between true and calculated slopes. Hammer *et al.* (1995) classified slopes into 5° classes and reported the percentage of cells in the matrix correctly classified, and correct to within one class.

Methods

3.1 GPS-based Ground Truth Data Collection

A Trimble GPS Pathfinder Pro XL system was used to collect ground truth location/elevation data. The system utilised a radio-link to a GPS base station service to deliver real-time differential positions with nominal sub-metre accuracy given a precision dilution of position (PDOP) of 4 and satellite elevation mask (SEM) of 15 degrees. Coordinate data from the GPS were recorded using the same coordinate system as the digital contours and topographic base data (the New Zealand Map Grid), allowing "way points" to be aligned as closely as possible with the 25 m grid of the DEMs interpolated from contours, given limitations imposed by trees and rock bluffs at a small percentage of sites.

Data was collected for two areas in the vicinity of Mt Vernon, on the Port Hills south of Christchurch (Fig 1). Area one included 400 points on a 25 m grid (500 m^2) for the north facing slopes below Mt Vernon, and area two 100 points in a 250 m^2 area over a rolling ridge crest north east of Mt Vernon. For the majority of the data collection PDOP values remained at 4 or better. However, in the deepest parts of the gully in the larger of the two areas, the steep terrain and limited horizon resulted in fewer satellite links, higher PDOPs, and lower positional accuracy. Data from the GPS were converted to a "ground truth DEM" simply by allocating each grid square the elevation value recorded at its centre.

3.2 NZLRI Slope Mapping

The study area on the Port Hills was mapped during the 1st edition phase of NZLRI mapping (Hunter, 1976), but has not been remapped to 2nd edition standards. To use only 1st edition data for an accuracy analysis with DEM slope data would not be a fair reflection on the whole NZLRI database, which contains, particularly in the North Island, substantial areas of 2nd edition mapping. To make an approximate assessment of the accuracy of slope mapping in the 2nd edition NZLRI, two scientists who were involved in both 1st and 2nd edition NZLRI mapping carried out a blind resurvey of the study area and surrounds (i.e., without exact knowledge of the location of the GPS survey) by interpreting aerial photographs and topographic contour data. Their slope maps (now referred to as 2nd edition NZLRI) were digitized, converted to 25 m raster format and compared with the ground truth slope maps. In addition, a detailed soil survey of the Port Hills (Trangmar, 1991) that included a classified slope attribute was converted to a 25 m resolution grid for comparison.

3.3 DEM Generation

Three 25 metre resolution DEMs were generated from three commonly used interpolators using Land Information New Zealand (LINZ) 20 m contours from the 1:50,000 topographic database. Two of the interpolators were from within ARC/INFO. They are referred to here as the ARCTIN and TOPOGRID methods. The ARCTIN method

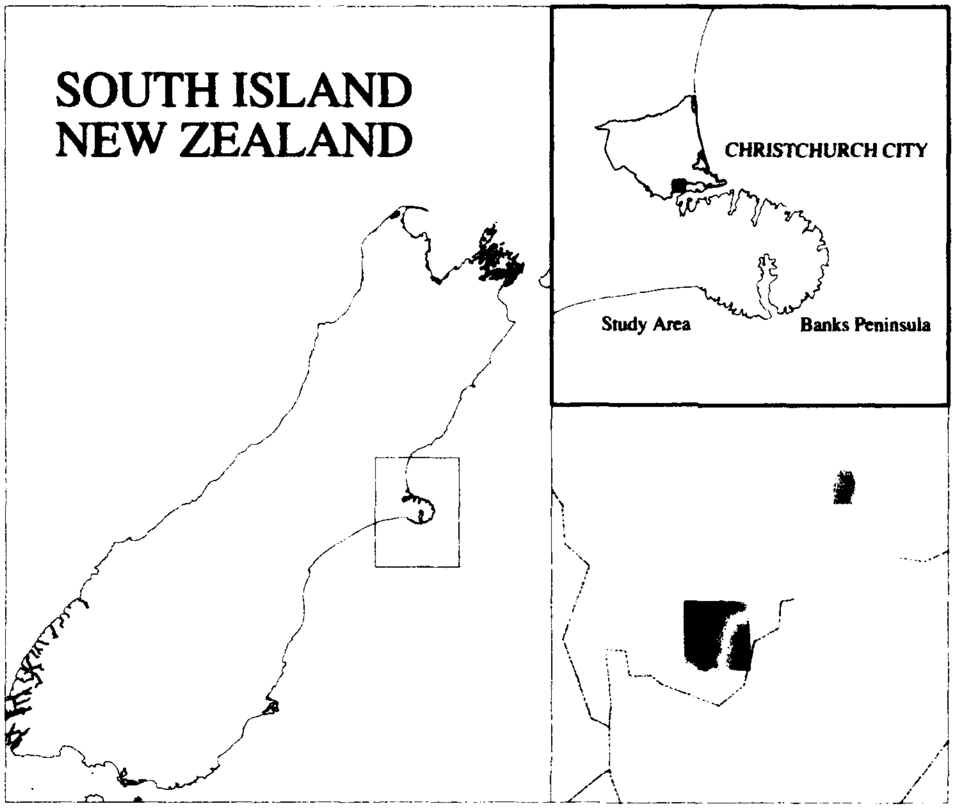


Figure 1: location map

uses CREATETIN to make an irregular triangular network (TIN), then uses the TINLATTICE command with the linear option to convert the TIN to a DEM. The TOPOGRID method is the ARC/INFO implementation of ANUDEM (Hutchinson, 1989). An interpolator developed in-house by Landcare Research, referred to here as the GILTRAP method (Giltrap, in prep.), was also used.

3.4 Slope Generation

Slope was calculated using the ARC/INFO GRID function SLOPE, which utilises a 3 x 3 window to calculate slope using the third-order finite difference method originally proposed by Horn (1981).

$$slope = \arctan\sqrt{\left[\left(\frac{\sigma z}{\sigma x}\right)^2 + \left(\frac{\sigma z}{\sigma y}\right)^2\right]} \quad (3)$$

where $\frac{\sigma z}{\sigma x} = \frac{((a+2d+g)-(c+2f+i))}{8 \times cell-resolution}$

and $\frac{\sigma z}{\sigma y} = \frac{((a+2b+c)-(g+2h+i))}{8 \times cell-resolution}$

with 3 x 3 cell notation as follows

a	b	c
d	e	f
g	h	i

Slope maps were generated from the DEM created by the ARCTIN method ("ARCTIN interpolated slope map") and from the DEM created from the GPS data ("GPS ground truth slope map").

	ARCTIN	GILTRAP	TOPOGRID
RMSE	5.77	7.76	7.94
mean error (_)	0.29	5.00	-4.06
S	5.76	5.85	6.80

Table 1: *Comparison of accuracy statistics for DEM surfaces generated from 20 m digital contours using different interpolation algorithms. These figures compare favourably with published USGS DEM standards for Level 1 DEMs for which 'a vertical RMSE of 7 meters or less is the desired accuracy standard', and 'an RMSE of 15 meters is the maximum permitted' USGS (1996).*

4. Results

4.1 Accuracy of Elevation Estimates

Table 1 presents the RMSE, _ and S statistics from a comparison of DEMs created by the ARCTIN, TOPOGRID, and GILTRAP methods with the ground truth DEM. The table shows that the ARCTIN-based DEM had the lowest RMSE, with the GILTRAP and TOPOGRID DEMs having similar RMSE statistics. The ARCTIN DEM also had the lowest _ and standard error statistic. The GILTRAP method generally over-estimates elevation by 5 metres, while the TOPOGRID method under-estimates elevation by a similar amount. The standard error statistic suggests that the ARCTIN and GILTRAP methods perform slightly better than the TOPOGRID method when interpolating from contours alone.

For all three methods these figures compare favourably with those quoted as standard for USGS DEMs (USGS, 1996) despite the use of trig stations and spot heights as ground truth points in USGS analyses. For example, level 1 USGS 7.5 minute DEMs must have an RMSE of less than 15 metres, and preferably less than 7 metres.

4.2 Accuracy of Slope Estimates

The classification matrix (table 2) illustrates the match between the ARCTIN interpolated slope map and the GPS ground truth slope map when both maps were classified into 5° classes. Some 36% of cells have the correct slope class assigned (±2.5°), while 83% of all cells are correct to within one slope class (±7.5°). This compares favourably with results reports from an analysis using a 30 m USGS DEM (Hammer, 1995).

In any calculation used for estimating slope from a DEM surface, elevation errors in the DEM surface are propagated through to the slope map. In this study maximum

ARCTIN Interpolated Slope Class	GPS Ground Truth Slope Class								
	Class	1	2	3	4	5	6	7	Total Cells
	0	12	4	1					17
	1	4	6	1					11
	2	1	9	5	1				16
	3	3	23	34	22	8	2	2	94
	4		13	35	69	61	6	2	186
	5	2	3	10	40	65	12	5	137
	6			2	12	26	10		50
	7			2	3	3	1	2	11
	8					2		1	3
	Total cells	22	58	90	147	165	31	12	525
	%correct	18.2	15.5	37.8	46.9	39.4	32.3	16.7	36.8
	% within one	77.3	65.5	82.2	89.1	92.1	74.2	25.0	83.4

Table 2: *Comparison of slopes derived from GPS ground truth survey and ARCTIN-based DTM. Of the 525 cells 37% are correctly classified (± 2.5°) and 83% are correct within one slope class (± 7.5°).*

elevation errors at any cell were found to be approximately ±20 metres. Within a 3 x 3 cell window this equated to a maximum possible error in height differential (see equation 3) of 40 metres over a distance of 50 metres (ie., twice the cell resolution) which results in a worst-case slope error of » 38°. Even at low error levels (±3 metres), the propagated error in slope is close to 5°, or within one 5° class of the true slope. The mean elevation error within the study area was ±5 metres which would result in a worst-case slope error of » 10°. The actual maximum error between the GPS ground truth slope map and the ARCTIN interpolated slope map was found to be » 22°, but errors of this magnitude are confined to a gully bottom where the GPS ground truth DEM is least reliable because of higher PDOP values. Elsewhere in the study area slope error rarely exceeded 10°, and was usually less than 5°.

4.3 Comparison with the NZLRI slope map

As the aim of this study is to determine whether a DEM derived slope map can improve on the NZLRI slope data, we have compared the NZLRI with the GPS ground truth slope map. All of the Mt Vernon study area falls within one 1st edition NZLRI polygon which is recorded as having F class slopes. This class includes slopes of between 26° and 35°. By reclassifying the GPS ground truth slope map using the NZLRI classification scheme, 'A' through 'G' (Water & Soil Division, 1969), only 7% of cells are correctly classified, while 35% are classified within one class of correct (ie., E or F slopes - between 21° and 35°).

The study area overlaps six polygons in the 2nd edition NZLRI which have slope classifications ranging from C (8-15°) to F (26 - 35°). When compared with the GPS ground slope map, 31% of cells are correctly classified, and 76% are correct within one class. Agreement between 2nd edition NZLRI and the ARCTIN interpolated slope maps shows good agreement (37% correct and 76% within one class) over the 3 km² area surrounding the GPS survey areas (Fig. 1).

Slope class data is also available from a 1:15000 scale soil survey of the Port Hills (Trangmar, 1991). Comparisons with the GPS ground truth slope map show that 43% of cells are correctly classified, and 90% are correct within one class. Over the 3 km² area surrounding the GPS survey areas the Port Hills soil survey map and the ARCTIN interpolated slope map also show good agreement (38% correct and 82% within one class).

5. Conclusions

5.1 GPS survey for Groundtruthing DEMs and Slope Maps

The GPS survey was useful as a rapid method for acquiring moderately accurate (±1m) locational data in the x, y and z dimensions. There was some difficulty in acquiring good data in some parts of the terrain studied because of a combination of poor satellite geometry during the middle part of the day and the degree to which the horizon was obscured when surveying in the bottom of the gully running through the study area. However, the ability to use the GPS to reproduce an independent 25 m grid of elevation values matching the contour-based DEMs, provided an objective method for estimating ground truth slopes over the same slope length as the DEM slopes. This method is still subject to errors, but they are more easily quantified than when comparing DEM slopes with those measured by clinometer over varying slope lengths.

5.2 DEM Accuracy

The ARCTIN and GILTRAP interpolation methods provide the most accurate DEMs, however the TOPOGRID method also gives a good surface with the added advantage of being hydrologically correct. However, the analysis suggests that the quality of input data from which the DEM is generated has a more significant effect on DEM quality than do the algorithms employed by the different methods tested.

All DEMs produced from LINZ 20 m contours would meet USGS standards (USGS, 1996) for a level 1 DEM. Twenty-five metres appears to be a practical resolution for DEMs to be used in conjunction with 1:50 000 scale data. Coarser resolutions will result in steadily increasing levels of error, while finer resolutions present substantial data storage and

manipulation problems for a database the size of the NZLRI.

5.3 Slope Accuracy

Slopes calculated from DEMs are subject to significant errors, even for DEMs with low *RMSE* and S statistics. While in the order of 70% of cells may have slope class correctly assigned to within one class of true slope (i.e., ±7.5°), the magnitude of potential DEM errors which could be propagated through the slope calculation strongly reinforces the need for DEMs to be supplied with some ground truth data and error statistics to quantify the accuracy of the data.

Because of the fractal nature of real hill slopes (i.e., variable at any scale) compared with the slope data derived from contour-based DEMs, which will only show variability at the scale differentiated by the contours, care must be taken in interpreting DEM slope data. The analysis above is suggestive of fuzzy sets, in that the DEM slope estimated for any cell may not exactly match the real slope on the ground at that point, but the relationship between the two may be represented as a membership function. In this case the shape of the membership function could be defined by the data in the classification matrix (table 2). This uncertainty in matching real slope to DEM slope has significance if DEM slope data is to be used to determine the extent of land with slopes exceeding some threshold. For example we might define erosion prone land as areas with slope greater than 15°. Clearly, for a proportion of cells in a DEM slope map with estimated slopes greater than 15°, real slope will be less than 15°. Similarly cells below the threshold might well have a real slope greater than 15°. Any analysis of DEM-based slope data must take this into account.

5.4 NZLRI slope classification comparison

The ARCTIN interpolated slope map (36% correct and 83% within one class) provided a significant improvement in accuracy for the study area over the 1st edition NZLRI data (7% and 35%). This differential is significantly less with simulated 2nd edition standards of NZLRI mapping (31% and 76%). The comparison with the detailed soil survey indicates that DEM-derived slope maps are approximately on a par with data collected in a field survey at 1:15 000 scale (43% and 90%). Slope maps derived from DEMs clearly give a significant improvement in resolution over traditional NZLRI mapping, as well as providing, for the first time, enough information to objectively estimate the magnitude of error in slope maps. These gains must be offset against the "no news is good news" perceptions of some data users, who may conclude that the DEM data is less reliable because the level of error is known, or even worse that the DEM data is 64% wrong because only 36% of cells are correctly classified.

Acknowledgements

David Giltrap and Stephen McNeill for their assistance in preparing DEMs. John Dymond, Stephen McNeill and Megan Ogle-Mannering for reviewing drafts of this paper. This research was supported by the Foundation for Research Science & Technology and research contract C09518.

References

Dymond, J. R., DeRose, R.C. & Trotter, C.M. (1992a). DTMs for Terrain Evaluation. *Geocarto International* 2:53-57.

Dymond, J. R., Stephens, P.R., Newsome, P. J. & Wilde, R.H. (1992b). Percentage vegetation cover of a degrading rangeland from SPOT. International Journal of Remote Sensing 13(11): 1999-2007.

Dymond, J. R. & Harmsworth, G. R. (1994). Towards automated land resource mapping using digital terrain models, ITC Journal 1994-2:129-138.

Dymond, J. R., DeRose, R.C. & Harmsworth, G. R. (1995a). Automated Mapping of Land Components from Digital Elevation Data. Earth Surface Processes and Landforms 20:131-137.

Dymond, J. R., Page, M. J. & Brown, L. J. (1995b). Updating NZLRI vegetation from satellite imagery. New Zealand Association of Resource Management Broadsheet July, 22-24.

Fryer, J. G., Chandler, J.H. & Cooper, M.A.R. (1994). On the Accuracy of Heighting from Aerial Photographs and Maps: implications to process models. Earth Surface Processes and Landforms 19:577-583.

Giltrap (in prep). Interpolating Digital Elevation Models from Contours.

Hammer, R.D. et al. (1995). Slope Class Maps from Soil Survey and Digital Elevation Models. Soil Science Society of America Journal 59:509-519.

Horn, B.K.P. (1981). Hill Shading and the Reflectance Map. Proceedings of the I.E.E.E., 69:14.

Hutchinson, M.F. (1989). A new method for gridding elevation and stream line data with automatic removal of pits. Journal Hydrology 106:211-232.

Hunter, G.G. (1976). New Zealand Land Resource Inventory Worksheet S84, scale 1:63 360, Water & Soil Division, Ministry of Works & Development, Wellington, New Zealand.

Li, Z. (1988). On the Measure of Digital Terrain Model Accuracy, Photogrammetric Record, 72(12): 873-877.

Monckton, C.G. (1994). An Investigation into the Spatial Structure of Error in Digital Elevation Data, in Worboys (Ed.), Innovations in GIS 1, Taylor & Francis pp 201-211.

Skidmore, A.K. (1989). A Comparison of Techniques for Calculating Gradient and Aspect from a Gridded Digital Elevation Model. International Journal of Geographical Information Services 3(4):323-334.

Trangmar, B.B.T. (1991). Soil Map of the Port Hills, Canterbury, New Zealand, scale 1:15 000, DSIR Land Resources Map Nos. 318, 319, DSIR Land Resources, Christchurch, New Zealand.

USGS (1996). Standards for Digital Elevation Models. http://ppc2.gis.univie.ac.at/software/dem/

Water & Soil Division (1969). Land Use Capability Survey Handbook. Ministry of Works & Development, Wellington, New Zealand, 139 pp.

Wilde, R.H. (1996). Updating the Vegetation Layer of the NZLRI National Database Using Satellite Imagery. New Zealand Geographer 52(1): 12-20.

The Nature and Management of Positional Relationships within Spatial Databases

David Lemon and Ewan Masters
School of Geomatic Engineering, The University of New South Wales
Sydney, Australia, 2052
ph 61 2 9385 4186
fax 61 2 9313 7493
email davidl@navigate.com.au
email e.masters@unsw.EDU.AU

Presented at the second annual conference of GeoComputation '97 & SIRC '97, University of Otago, New Zealand, 26-29 August 1997

Abstract

This paper gives a description of the results of an investigation into the nature and use of positional relationships in GIS. The positional relationships between features in spatial databases used in local government have been analysed to determine a set of rules that enable the spatial integrity of databases to be maintained during update and upgrade procedures.

1. Introduction.

The increasing use of Geographical Information Systems (GIS) throughout the community has prompted people to investigate the way in which positional information is managed. Whilst there are a number of methods for storing and representing spatial information, very few commercial GIS packages include the ability to store and manage the relationship information that is used to collect or create this data.

The existence of these relationships has been recognised by a number of authors (Kjerne and Dueker 1986, Corson-Rikert 1988, Driessen and Zwart 1989, Hebblethwaite 1989). They have been referred to as 'associativity' (Hesse 1991, Wan 1993, Baker and Paxton 1994), 'relativity' (Hadjiraftis and Jones 1991, O'Dempsey and Moorhead 1991), 'graphical data dependencies' (Unkles 1992), 'vertical topology' (Blackburn 1994, Lemon 1995) and 'positional relationships' (Lemon 1997). For this paper, the term ***'positional relationship'*** will be used as it is felt that, unlike other terms, it helps to describe the nature of these relationships.

Here a ***positional relationship*** is defined as:

> *A relationship between spatial features that has been used, primarily, to determine the real world position of one of those features with respect to the other features, during initial data capture.*

For example, if, in order to determine the position of a sewer entrance, physical measurements are made to nearby fence corners, then positional relationships exist between the entrance and each of these corners. Further, if a particular administrative boundary (eg. an electoral boundary) is defined as being coincident with some other boundary (eg. a Local Government Boundary), then a positional relationship exists between these two boundaries.

The use of positional relationships to determine the positions of features in the real world pre-dates GIS technology. All coordinates are relative to some real world feature, be it the monuments that form a survey control network, or any other real world features. Traditionally, positional relationships have been used by cartographers to position features on maps. However, once the position of a particular feature has been determined and that feature placed on the map, the relationship information is discarded. The position of the feature is then only represented by a set of coordinates, which are implied on a

paper map, or actually stored for a digital map. The relationship information is therefore discarded.

The removal of positional relationship information creates a problem for the management of features in spatial data sets that undergo change. It is quite possible that the positions of features in other data sets may have been determined using positional relationships with respect to these moved features and that these relationships are still relevant and important to the spatial integrity of the database. However, because the positional relationship information has been discarded it is very difficult to maintain the positional integrity of the various themes of data.

The problem of positional changes affecting the spatial integrity of other related data sets has been recognised for a long time. However the only solutions developed seem to have concentrated on the method of maintaining specific relationships without attempting to define the nature of positional relationships or define the criteria for whether positional relationships need to be maintained or not.

The purpose of this paper is to outline the results of an investigation into the nature of positional relationships and the role this nature plays in determining how the processes of positional change affect different relationships. These results can be used to develop a system for the comprehensive management of positional relationships within any GIS and therefore maintain the integrity of spatial databases under the various scenarios where features in a database have their coordinates changed.

2. *The Process*

The investigation involved a number of steps. Firstly, it was necessary to study the different types of spatial information being collected and used. In order to make the scope of this assignment manageable, spatial data used in local government was investigated. This application area was chosen as a number of studies have shown that the use of GIS technology within local governments is very widespread (Craig 1994, Masser and Craglia 1995, Marr and Benwell 1996) and the data and data collection techniques used by these organisations is extremely diverse (Lemon 1995). It can therefore be safely assumed that the number and diversity of positional relationships used by these organisations will be such that the results of the investigation can be readily be applied to other areas of GIS. The actual datasets were determined by surveying local councils about the creation, maintenance and application of their spatial data (Lemon, 1995).

Secondly a systematic investigation into the positional relationships between features in each of these data sets was undertaken. In particular, this investigation studied which types of spatial features are related, how they are related and why they are related. The results of this investigation are described in Section 3.

The next step involved studying the different types of positional changes that occur within spatial databases. It was necessary to look at, not only how and when these changes occur, but also their affect upon positional relationships. Section 4 details the results of this study.

Finally, using the results of these investigations, a set of requirements and a rules base for the management of positional relationships were developed. These will be given in Section 5.

3. *The Nature of Positional Relationships*

The investigation into the types and use of positional relationships revealed that it was necessary and possible to classify relationships in order to define criteria for maintaining them. Firstly, it was found that in any relationship there may exist one or more base, or 'Master', features but only one related, or 'Slave' feature such that the position of a Slave feature is dependent, through the positional relationship, upon the position(s) of its Master feature(s). Secondly, the many different forms of relationship were analysed and categorised (eg. Bearing and Distance, Relative Position, Offset, etc.). Finally, two distinct categories or class of relationship, based upon the purpose for which the relationship exists, were identified.

3.1 The Place in a Positional Relationship

Place in Positional Relationship	
Master	Slave

It is convenient to define the term **place** in the positional relationship as being either **master** or **slave**. That is, a feature can be a master feature (or base feature), from which the positions of slave features are determined or, the position of a slave feature is determined with respect to a master feature. The place of a feature within a positional relationship affects how a positional change will affect that feature.

3.2 The Form of a Positional Relationship.

In general, positional relationships are used to determine the positions of point and line features. They can be either specific measurements, such as a distance, or they can give only a general indication of the position of a feature without allowing for the calculation of an actual coordinate. This situation occurs when one feature is placed 'relative' to some other feature.

Positional relationships can take a great many forms. However, after thorough analysis of data acquisition techniques, a set of fundamental forms that can be used to describe any positional relationship can be defined as in Table 1.

A number of these forms cannot be used alone to determine a unique position for a Slave feature. However, in combination they can be used to describe most if not all positional relationships. For example, a 'bearing and distance' relationship can be described using a **distance** relationship and an **angle to slave relationship**.

3.3 The Class of a Positional Relationship

Another categorisation method is by the purpose for which the relationship exists. This method of categorisation is defined here as the **'class'** of a relationship.

Two distinct classes of positional relationships are used.

1.	Distance - the Slave point lies on a line which is a circle of known radius centred on one Master point;
2.	Angle to Slave - the Slave point lies on a straight line which intersects, at a known angle, another straight line between two Master points at one of these Master points;
3.	Angle at Slave - the Slave point lies on a line such that, at all times, the angle between two straight lines from that point to two Master points remains constant;
4.	Distance Along a Line - the Slave point lies on a Master line at a known distance along that line from a Master point;
5.	Point Offset - the Slave point lies on a line which is parallel to, and a known perpendicular distance, from a Master line;
6.	Intersection - the Slave point is defined by one of the intersections of two Master lines;
7.	Line Offset - the Slave line is defined as being a line parallel, to and at a known perpendicular distance, from a Master line;
8.	Relative Position - the Slave feature (point or line) is in a position relative to a Master feature (line or polygon); and
9.	Models - the positions of features in the Slave data set have been calculated using a mathematical model and the information in one or more Master data sets.

Table 1 Positional Relationship Forms (Lemon 1997)

They are:

1. a **Measured Relationship** where the position of a feature at a particular time, has been determined using either measurements to other features or by being placed relative to some other feature.

2. a **Defined Relationship** where the position of one feature is defined for all time by its relationship to another feature.

Class of Positional Relationship	
Measured	Defined

The differences between these two classes, on the surface, appear quite small. However, the importance of this method for categorising positional relationships is that a relationship's class is one of the major factors contributing to how that relationship will be affected by a positional change to a feature involved in the relationship. The effects of the different types of positional change on different classes of positional relationship will be discussed in the following section.

In the case of a measured relationship, the relationship only exists for the purpose of determining the position of the Slave feature with respect to the Master features at the time of measurement. Driessen and Zwart (1989) refer to these features as being 'statistically independent' as, *whilst they are related in the GIS, in the real world they are very much independent.* Examples of this relationship class occur in the *utility industry,* where the position of certain utilities is determined by measurements to some base feature, usually the parcel boundaries.

In the case of a defined relationship, however, the purpose for the positional relationship is quite different. In these relationships, the position of the Slave feature in the real *world, is defined by its relationship to its Master feature(s)* until this definition is changed. At any time, in order to know where the Slave feature is, it is necessary to know the position of the Master feature. The features are referred to as being 'statistically dependent' by Driessen and Zwart (1989), as the position of the Slave feature is very much dependant upon the position of its Master. Examples of defined relationships are associated with the many administrative boundaries used by the various levels of government. One boundary may be defined as being coincident with a particular base feature (eg. a road centreline).

4. *Positional Change in a Spatial Database.*

It is unwise and incorrect to assume that the position of a spatial feature will remain unchanged for the life of a spatial database. Positional Changes take a number of forms and occur for a number of reasons. In each case however, it is necessary to ensure that the database is a true representation of the real world and in order to maintain the integrity of the spatial database it is necessary to make *these changes within the GIS.*

Analysis of the types of positional change that can occur reveals that it is possible to categorise positional changes in a spatial database into two types:

1. an **update** where the position of the feature has changed in the real world (see Masters 1988, Baker and Paxton 1994). Such a change will occur if a feature is moved, created or ceases to exist, in the real world. In each case, it is necessary to reflect this change in the spatial database in order to maintain its currency. If this is not done the *information* in the GIS will fail to represent reality.

2. an **upgrade** where new information about the original position of the feature has been obtained (Masters 1988, Baker and Paxton 1994). In this case, the position of the feature, in reality, has not changed at all. Rather, the set of coordinates used to represent the position of the feature in the GIS has been replaced by a new set of coordinates. This new *information* may have a number of sources and may, or may not, be more accurate than the original information, depending upon its source.

Type of Positional Change	
Update	Upgrade

Both types of positional change occur in some spatial databases regularly. In both cases the change will result in a change *to the positional representation* of features within

the GIS. In turn, any positional relationships that may have existed between these features and other features will be affected in some way. It is therefore important to understand the differences between the affect of updates and upgrades on positional relationships.

5. Positional Relationships and Positional Changes.

5.1 Updating and Positional Relationships

An example of an update is shown in Figure 1. In this example, the Parcel boundary, to which the Water Main and the Building Line are related, is updated such that the distance between the lines a-a' and b-b' is widened by 5m.

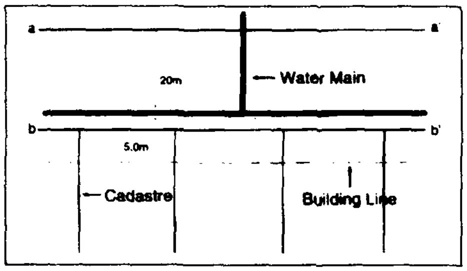


Figure 1
Before the Update. The parcel boundary (full line) is the Master to which both the water main (bold) and the building line (dashed) have been related. The water main is related to the parcel boundary by measurements, whilst the position of the building line is defined as being 5.0m offset from the parcel boundary

The effect of an update on a positional relationship is very much dependant upon the class of that relationship. In the case of measured relationships the positions of the features involved are independent. Hence, a change to one of these features will not affect the positions of other features. Therefore, after an update a measured positional relationship is no longer valid.

Thus, in the example in Figure 1, the position of the water main had originally been determined from two relationships to points on the parcel boundary. These are measured positional relationships hence the update will have no effect upon the position of the Water Main. The result of the update upon these relationships, is to render them invalid.

For features involved in defined relationships the result of an update to the Master feature(s) will be dependant upon exactly what happened to that feature in reality. In the case where a Master feature is moved, it may actually be necessary to perform a similar move to the Slave feature as the relationship may still be valid. It is therefore necessary to ensure that the relationship is maintained.

For updates where the Master feature has ceased to exist, however, two possibilities exist. Firstly, the relationship may become invalid and the related feature will continue to exist independent of any relationship. Secondly, it may be necessary to define a new relationship for the position of the Slave feature.

In the example in Figure 1, the Building Line is defined as being offset 5.0m from the Parcel boundary. The update, has caused the parcel boundary to be moved. In order to ensure the spatial integrity of the GIS, it is necessary to perform a similar move to the Building Line in order to maintain the offset requirement.

In the examples used so far, the updated feature has been a Master feature. It is also necessary to look at the effects

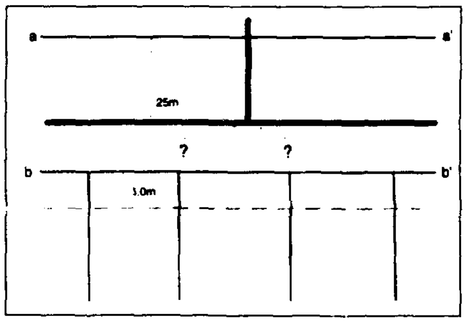


Figure 2
After the Update. Due to the different relationship classes, the update to the parcel boundary has affected the two related features in different ways. The building line has also been updated in order to maintain the 5.0m offset requirement. The water main however, does not require updating as its relationship to the parcel boundary was used simply to determine its original position.

upon a positional relationship in cases where the update occurs to a Slave feature. Here, the effect of the update will be the same as that for updated master features. That is, the relationship will cease to exist. Similarly, in cases where an updated slave feature is involved in a defined relationship, the positional relationship will also become invalid, as the act of updating the feature implies that the definition of the position of that feature has been changed. That is, the slave feature no longer exists or its position has been redefined with respect to some other feature.

5.2 Upgrading and Positional Changes

An upgrade implies that new information about the position of a feature has been obtained or, in some cases, calculated, despite the fact that the feature has not moved in reality. It can be seen that as the upgraded feature has not moved, any positional relationships (both defined and measured) that existed between it and other features are still relevant. Thus, in order to ensure the positional integrity of these related features, they should also be upgraded such that the affected positional relationship is maintained. This result is quite different to that for updates.

For both measured and defined relationships, an upgrade to a master feature will require the slave features to also be upgraded using the form of the relationship. An upgraded position for a slave feature implies that new information about its original position has been determined. Thus, given that the positional relationship was only used to determine this original position, the new information overrides the relationship which becomes invalid.

The effect of an upgrade to a slave feature upon a defined relationship actually causes a conflict of information. In this case, given that the position of the slave feature was originally defined via a positional relationship form, theoretically, it should not be possible to determine new information about the original position of that feature. Therefore, an upgrade to the position of the slave feature does not imply that the definition was incorrect, rather it implies that the original position of the master feature was incorrect. However, to enforce the relationship and reposition the master feature would imply that its position was dependent upon the position of the slave feature. Whilst, in reality, this could be possible, for consistency it should not be allowed, as it contradicts the definition of a positional relationship.

6. *The Result*

There have been three possible solutions proposed for maintaining relationships (Hebblethwaite 1989). They are the Transformation method, the Database method and Object-Orientation. Each of these techniques have been discussed elsewhere and will not be discussed here (see Wan 1993, Lemon 1995, Lemon 1997). The findings of this investigation into the affects of different forms of positional change on different positional relationships show that whilst the maintenance of a relationship may be required in some situations it will not be required in all situations. Also, techniques that have been developed to maintain a specific "form" of relationship cannot possibly manage all positional relationships in a GIS. It is therefore obvious that only the database and object-orientation techniques can be applied to fulfilling the requirements of a positional relationship management system that will truly maintain the spatial integrity of a database under conditions of updating and upgrading. This problem is discussed in more detail in Lemon (1997).

In order to manage the relationships between spatial data it is necessary to develop a method which is able to determine the type of positional change that has occurred, the class of the affected positional relationship and the place the affected feature holds in that relationship. Using this information a set of rules based on the above findings can be developed to determine the required action to take with respect to other features in the affected relationship. The effects of different types of positional change upon different types of positional relationship are summarised in Table 2.

A system for managing positional relationships must be able to perform a number of functions. They are:

- Detect all positional changes and determine their category and type (update or upgrade).

Type of Change	Feature Place	Relationship Class	Affect on Positional Relationship
Update -	Master	Measured	Relationship becomes invalid.
Moved		Defined	Relationship may remain valid
	Slave	Measured	Relationship becomes invalid.
		Defined	Relationship becomes invalid
Update -	Master	Measured	Relationship becomes invalid.
Deleted		Defined	A new Relationship should be defined.
	Slave	Measured	Relationship becomes invalid.
		Defined	Relationship becomes invalid.
Upgrade	Master	Measured	Relationship remains valid.
		Defined	Relationship remains valid.
	Slave	Measured	Relationship becomes invalid.
		Defined	Conflict of Positional Information. User must decide.

Table 2 The Effect of Positional Changes on Positional Relationships.

- Identify any positional relationships affected by these changes.
- Determine the correct action to take with respect to that relationship.
- Perform that action.

One of three possible 'actions' can be taken when managing positional relationships. The relationship can be maintained, it can be extinguished, or a new relationship can be defined. However there is one situation where a different action may be required. This occurs when an upgrade to a Slave feature has occurred as a direct result of its relationship to its Master feature(s) having being maintained. In these cases the positional relationship remains valid and the Master and Slave features should be in their correct relative positions. Hence the necessary action is to 'do nothing'.

The above case shows that an essential part of determining the required action, is the ability to determine the current status of the features involved in the relationship. That is, it is necessary to be able to determine whether these features are in their correct positions with respect to the relationship. In order to perform this function, it is necessary to be able to reconstruct specific relationships. Table 3 shows the action that should be taken for each case of a positional change occurring to features in a spatial database.

7. Conclusion

The management of positional relationships in GIS is gaining increasing recognition as being very important for the GIS industry. This is due to the following facts:

- Much of the spatial data used by these organisations contains features which are closely related;
- The features in some of these data sets are undergoing constant positional change;
- Some of these changes are having the effect of degrading the spatial integrity of data sets containing features which are related to these updated or upgraded features;
- If these relationships are not managed correctly, the spatial integrity of these data sets will, over time, become so severely degraded as to make them useless.

Much of the previous work in this area seems to have concentrated on implementing methods to maintain the "form" of relationships without analysing in detail the problem that needs to be solved. This study has therefore in-

Change	Place	Class	Rel. Status	Affect on Relationship
Update -	Master	Measured	*Not Applicable*	Extinguish Relationship.
Moved		Defined	*n/a*	Maintain Relationship.
	Slave	Measured	*n/a*	Extinguish Relationship.
		Defined	*n/a*	Extinguish Relationship.
Update -	Master	Measured	*n/a*	Extinguish Relationship.
Deleted		Defined	*n/a*	Extinguish Relationship and
				Define New Relationship.
	Slave	Measured	*n/a*	Extinguish Relationship.
		Defined	*n/a*	Extinguish Relationship.
Upgrade	Master	Measured	*n/a*	Maintain Relationship.
		Defined	*n/a*	Maintain Relationship.
	Slave	Measured	Correct	Do Nothing
			Incorrect	Extinguish Relationship.
		Defined	Correct	Do Nothing
			Incorrect	Conflict of Positional Information. User must decide.

Table 3 The Action Required to Manage Positional Relationships. The required action may be dependent upon the type of positional change (change), the feature place, the relationship class (class) and the current status of the relationship (rel. Status).

vestigated the nature of positional relationships and positional change in spatial databases commonly used in local government. However, these results should be generally applicable to any spatial database. The many complex relationships between features have been classified into simple categories that can be used to build a set of rules for how and when to maintain relationships between features in spatial databases. The implementation of a positional relationship management system would require appropriate positional relationship data to be stored at the feature level with other spatial data. Once this type of data is stored it will then be possible, using either a database approach, or object-oriented approach to determine the correct action that should apply to positionally related features in a database and mainatin the integrity of spatial databases.

8. Acknowledgements

David Lemon was supported by an Australian Postgraduate Research Award (Industry) for this work along with sponsorship from GenasysII. We gratefully acknowledge the many contributions John Blackburn has made to this work.

9. References

Baker A. and Paxton I., 1994, *The Associativity Problems in the Utility Environment: A Progress Report on the Problem Solving Process*, Proceedings: AURISA 94. Sydney November 1994, p521-527.

Blackburn J., 1994, *GENASYS Vertical Topology*, Genasys Technical Memorandum, Genasys II Pty. Ltd. September, 1994.

Craig W., 1994. *The Rising Tide of GIS in Minnesota*, The Journal of the Urban and Regional Information Systems Association. 6(1), 75-80.

Corson-Rikert J., 1988. *Updating Multilayer vector databases*, Seminar on Resources Planning Management and Monitoring, IES Report 132, University of Wisconson-Madison, 147-159.

Driessen R. and Zwart P., 1989. *A Model for Representing Cultural Parcels in a GIS*, Proceedings: URPIS 17, Perth, Australia, p443-451.

Hadjiraftis A. and Jones A., 1991. *The Reconciliation of DCDB Dependent Spatial Data Sets with DCDB Upgrades*, Proceedings: Symposium on Spatial Database Accuracy, Department of Surveying and Land Information, The University of Melbourne, Melbourne. p95-112.

Hebblethwaite D., 1989. *Concepts for Coping with a Shifting Cadastral Model*, Australian Surveyor, 34(5), p486-493.

Hesse W., 1991. *Object-oriented Concepts for Land and Geographic Information Systems*, University of Melbourne, PhD Thesis. 150pp.

Kjerne D. and Dueker K., 1986. *Modelling Cadastral Spatial Relationships Using an Object-Oriented Language*, Proceedings: Second International Symposium on Spatial Data Handling, Seattle, Washington. p142-157.

Lemon D., 1995. *Vertical Topology - An Investigation into the Requirements of Local Government*, Proceedings: Aurisa 95, Melbourne Australia, November 1995, p378-390.

Lemon D., 1997. The Nature and Management of Positional Relationships in a Local Government Spatial Database, PhD Thesis, The University of New South Wales, Submitted: May 1997.

Marr J. and Benwell G. 1996. *The Evolution of the Use of GIS in New Zealand Local Government*, The New Zealand Surveyor 286, 30-35.

Masser I. and Craglia M., 1995. *The Diffusion of GIS in Local Government in Europe*, Proceedings: AURISA 95, Melbourne, Australia Nov 1995, 292-307.

Masters E., 1988. *Upgrading a Digital Graphic Database*, Proceedings: AURISA-URPIS 88

O'Dempsey T. and Moorhead C., 1991. *A Solution to the Problem of Relativity*, Proceedings: OZRI 5, Queensland, Australia.

Unkles H., 1992. *Owning the Solution Provides the Power*, Proceedings: AURISA 92, Gold Coast, Australia November 1992. p158-163.

Wan W., 1993. *Managing Associativity in Land Information Systems*, University of Melbourne, Masters Thesis. 91pp.

Outside-In, Inside-Out: Two Methods of Generating Spatial Certainty Maps

Kim E. Lowell
Centre de recherche en géomatique
Pav. Casault, Université Laval
Ste-Foy, Québec G1K 7P4
CANADA
(418) 656-7998
FAX: (418) 656-7411
Internet: Kim. Lowell@scg.ulaval.ca

Presented at the second annual conference of GeoComputation '97 & SIRC '97, University of Otago, New Zealand, 26-29 August 1997

Abstract

Problems with standard Boolean maps produced through subjective interpretation of a phenomenon (i.e., forest or soils maps) are discussed and two alternatives based on spatial certainty are presented and discussed. One requires multiple interpretations of the phenomenon in order to construct a library of spatial uncertainties; such an Uncertainty Library can be used subsequently to estimate error across cartographic boundaries. The other requires interpreters/cartographers to identify only those map elements which are "100% certain." A spatial interpolation algorithm is subsequently applied to this information to "fill in the gaps" with certainty information for each map type. Both methods have advantages and disadvantages relative to standard Boolean maps and also to each other. These are discussed in general terms and also through the presentation of specific examples. It is concluded that though uncertainty-based cartographic representations provide more flexibility than do conventional Boolean maps, the construction of the former is not without its problems either.

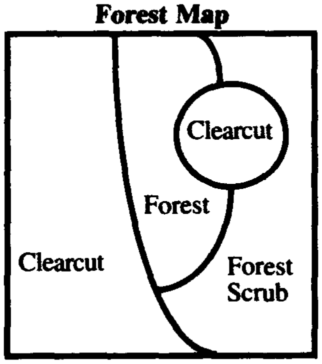
1. Introduction

In recent years, a number of researchers and practitioners have become interested in maps showing spatial certainty/uncertainty. These maps are often conceptualized as showing fuzzy membership values or probabilities for a given map class (e.g., Hall *et al.*, 1992, Lowell, 1994). One way to produce such maps is to obtain a standard thematic map of the variable of interest and use available information and/or make assumptions about the magnitude and nature of the error inherent in the variable mapped. The information on the Boolean map may then be "perturbed" stochastically to produce a fuzzy map of the variable under study (e.g., Fisher, 1992, Goodchild *et al.*, 1992). In this process, it is often the case that one considers only the cartographic type/value that was originally mapped at a given location rather than also considering additional spatial information that cartographers or others familiar with the variable mapped possess mentally. For example, one might perturb the map type "Forest" using various assumptions about the certainty of this type, but one might not consider whether or not a given "Forest" polygon is surrounded by "Lake" on one side or "Forest Scrub" on another. Yet these two types as neighbors of a Forest polygon imply very different things about the certainty within a polygon labeled "Forest." Hence it would seem to be useful to consider the characteristics of a map type at various places within a given polygon — e.g., close to/far from a boundary — and/or to consider this relative to an adjacent polygon of a given type — e.g., a Forest/Lake boundary should be treated differently than a Forest/Forest Scrub boundary. This also suggests two valid, yet seemingly contradictory approaches to the development of certainty or fuzzy maps.

In one, a single map and a "library of uncertainties" showing the certainty associated with a boundary of a given type are assumed to be available (Fig. 1). Edwards and Lowell (1996) have demonstrated how such an Uncertainty Library may be developed from multiple cartographic interpretations of a given phenomenon. However, such a process has some important limits for the work described; these will be discussed subsequently. For the moment, assume that an Uncertainty Library is available that is known to be applicable to a recently constructed Boolean — i.e., conventional thematic — map for a given phenomenon. The Uncertainty Library consists of the standard deviation for the true location of a boundary line separating any two types that may appear on the map. Given this information, it would seem to be a relatively simple matter to construct a certainty map from the available Boolean map: one identifies the cartographic types on either side of the boundary and applies the appropriate standard deviation. However, it will be demonstrated that the process is more complex than it appears here.

In the other approach, what is considered "a map" is radically different from a conventional Boolean map. Generally, one considers a map to be a complete coverage of a surface — one in which every location has a "value" relative to the variable being mapped. (Note that even a collection of points may be viewed this way ultimately since interpolation is usually used to assign a value to all gaps between points before the map is employed in any analysis.) However, in some disciplines such as natural resources, a standard Boolean thematic map represents a considerable loss of information concerning spatial uncertainty. For example, in the production of forest type maps for Quebec (Ministry of Natural Resources, 1995), aerial photographs are interpreted subjectively by trained human photo-interpreters. The author's experience has shown that in interpreting a photograph, an interpreter works from a "definite" object or area — e.g., a lake, a clearcut — and proceeds to less certain features. At various times the photo-interpreter places a boundary line because of 1)the actual recognition of a definite boundary or dividing line (with types on either side not necessarily being known), 2)the necessity to separate two regions of clearly different types for which the boundary location is not known exactly, or 3)due to the recognition of an actual closed polygon of a given cartographic type (Fig. 2). Note that in the traditional method a photo-interpreter does not always see closed polygons, but is forced to produce them nonetheless. The result is a Boolean map for which one must try to infer certainty from a subjective knowledge of the phenomenon being mapped. If photo-interpreters were permitted to produce an interpretation showing only those features having "100% certainty," it would be possible to

Figure 1. Example of map and associated Uncertainty Library. Library shows one standard deviation for location. (Representation inspired by R. Aspinall, pers. comm.)



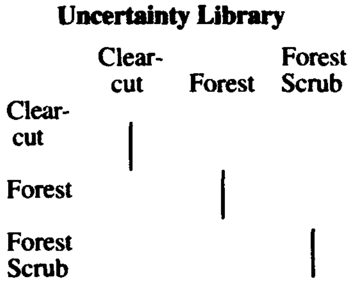
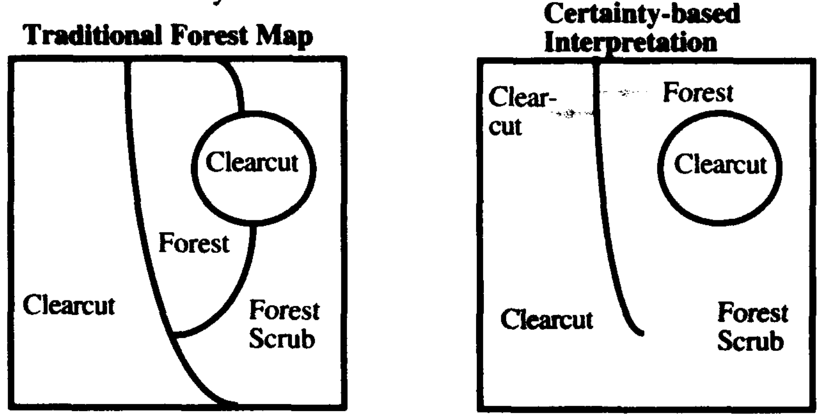


Figure 2. Traditional polygon map and alternative certainty-based interpretation. Regions and boundaries on certainty-based interpretation are "100% certain" whereas polygons on the traditional map have variable and unknown certainty.



derive a certainty map using spatial interpolation. In doing so, instead of inferring uncertainty, one would have explicit information and there would presumably be more consistence among interpretations. However, this process is also not as straightforward as it would seem.

At this juncture, the primary point being made is that, for the production of certainty maps, we have two possible alternatives to map perturbations and its accompanying assumptions. However, these alternatives also are subject to certain difficulties and assumptions. Moreover, the two would seem to be somewhat contradictory. One employs the boundary as its basic unit and works towards the center of polygons. That is, the high certainty implicitly associated with polygon cores is derived from observations at polygon boundaries. In the second, it is conceivable that only polygon cores will be identified by a photo-interpreter as being "100% certain." Thus the low certainty at boundaries is derived from the high certainty at polygon cores. Are the two methods compatible? What are the problems associated with each? Are there any particular benefits of one over the other?

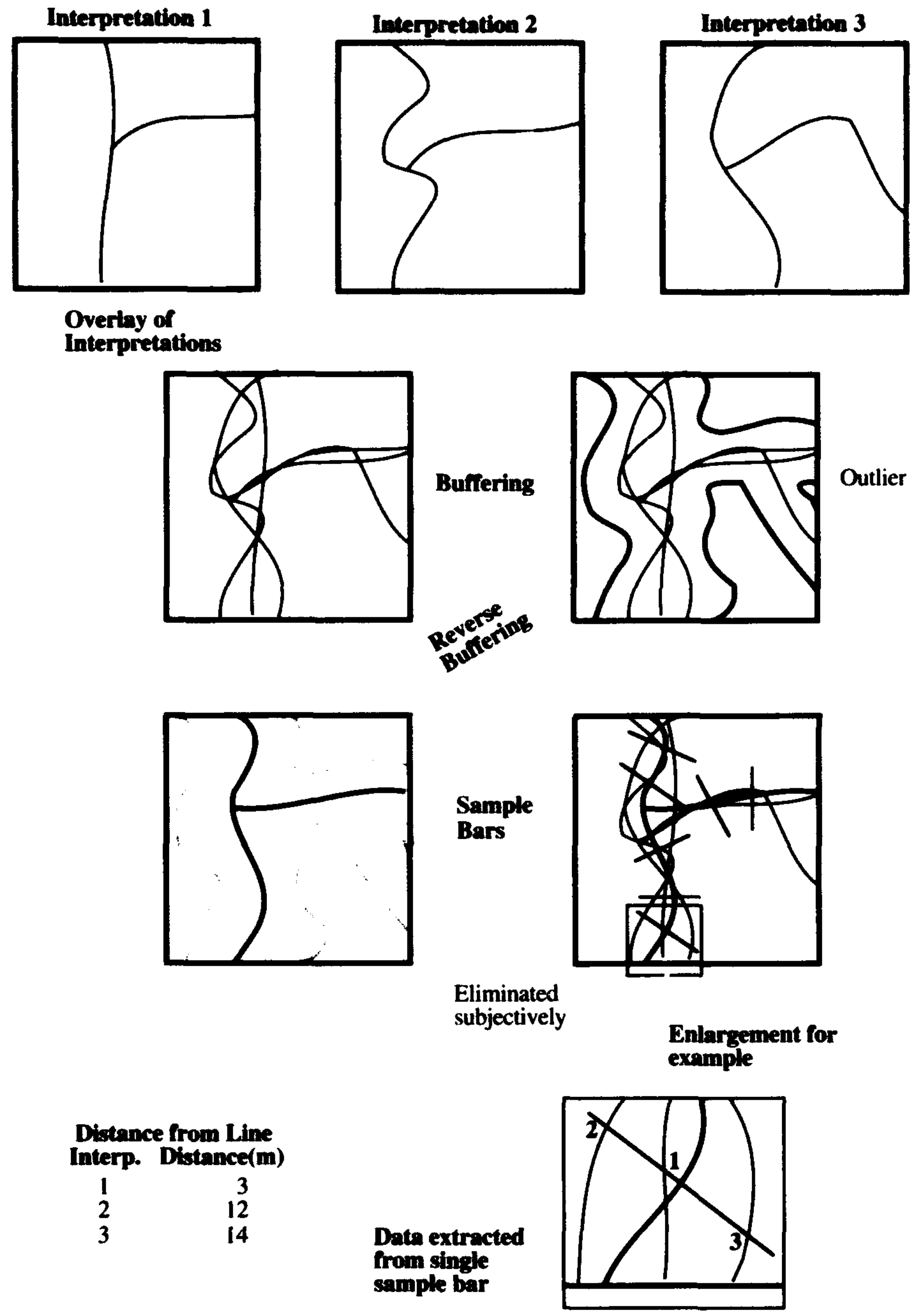
The purpose of this paper is to respond to these and similar questions and also to explore the two methods in greater detail. This includes not just computational aspects regarding the two, but also user considerations including data collection and organization.

2. *Method 1: Boundary-based Certainty Maps: Outside-In*

2.1 Data collection

If one is to use a single, conventional thematic map to generate a certainty-based map, knowledge about the associated error must be available *a priori*. In this paper, it is assumed that the form of this knowledge is a standard deviation on the location of a boundary of a given type (Fig. 1). In effect, this means that one must fill a k-by-k matrix (in which k is the number of map classes) with a standard deviation or other measure of boundary uncertainty. Put another way, we need to know the locational uncertainty of Forest/Clearcut boundaries, Forest/Forest Scrub boundaries, and Clearcut/Forest Scrub boundaries (assuming k=3 in this case). One of the most straightforward ways to obtain this information is through multiple interpretations of the same phenomenon. The method to be described was developed and described by Aubert (1995) and is presented schematically in Figure 3.

Figure 3. Method for obtaining Uncertainty Library from multiple interpretations.



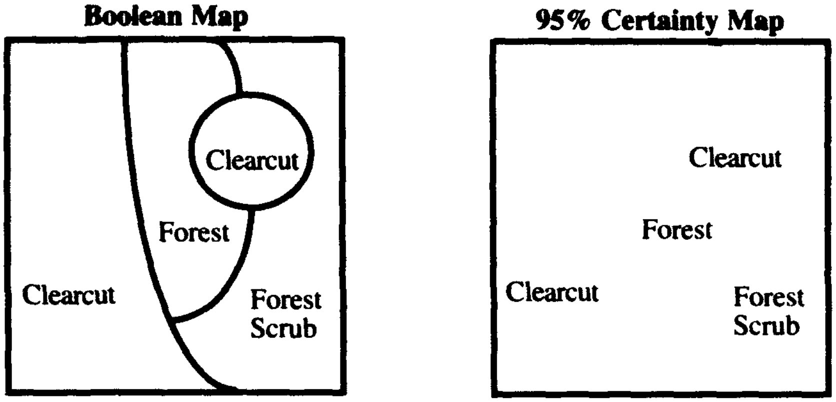
To develop a library of spatial uncertainties, one starts with multiple interpretations of the same phenomenon. As an overall goal, it is desired to use these interpretations to develop a map of "truth" by overlaying the boundaries, identifying the mean location for each boundary, and subsequently quantifying the error around each of these mean boundaries. Suppose for the purposes of explanation, three such interpretations are available (Fig. 3). The boundaries of these are overlaid and a buffering operation around all boundaries is performed. In doing so, the size of the buffer selected will strongly affect the results. This is because all lines within the buffering distance of each other will be bundled together and considered to represent the same "true" line. "Too large" a buffer will cause more than three lines to be bundled into one boundary — something that is clearly impossible if one has only three interpretations; a buffer that is "too small" will cause relatively few boundaries to be bundled together. Once the buffer size has been selected and the buffering operation performed, the outer limits of each buffer are retained and outliers removed manually. Note that this requires a subjective judgment on which lines are outliers (middle of Fig. 3) and sometimes causes less than *n* interpretations to define a line. A reverse buffering operation is then performed to identify the mean location — assumed to be the "true" line location — for each boundary. This mean/true boundary location is then overlaid on the original interpretations and a series of sample bars placed along the mean line. The bars are the size of the original buffer and are placed at selected intervals along the mean location; these sample bars are placed perpendicular to the mean/true line and are spaced far enough apart to avoid the effects of spatial autocorrelation. The distance each line on an original interpretation is from the mean/true line for each sample bar is determined and the mean and standard deviation of these calculated. These are then summarized by boundary type — i.e., all the Forest/Lake boundaries summarized together regardless of their location on the map.

2.2 Treatment and use of information

The method described provides an estimate of the error associated with each type of map boundary line and provides for the development of an Uncertainty Library. Moreover, it provides a means to test if there is a systematic bias in interpretation. For example, it might be the case that in looking for a Forest/Lake boundary on an aerial photograph, there is a consistent tendency to place the separating line towards/away from the Forest. This might be caused by shadows and/or the interpreter's eye consistently being drawn toward/away from water, and/or other factors. This also highlights another use of the methodology developed — one can test the nature of the distribution of the error across the mean location. This was done in the original study (Aubert, 1995); there was no evidence to reject the null hypothesis of the error across the mean line location being distributed according to a Gaussian distribution.

Knowing that the information sought — i.e., a fully populated Uncertainty Library — can be obtained, attention turns to its use. It is assumed here that the Uncertainty Library has been compiled in such a manner as to be robust enough to be used for an area for which an uncertainty map is to be developed. A single Boolean map is produced from a single photo-interpretation of the area, and one may now ask questions such as "Show all the areas which have a probability *p* of being Type A." For example, asking for the 50% confidence interval for all map types will produce the Boolean map itself since, effectively, each boundary represents the point at which there is a 50-50 chance of being the type on either side of the boundary. Similarly, one may want the 95% confidence interval on Clearcuts. A Gaussian distribution can be generated around all Clearcut boundaries using the Uncertainty Library and the point at which 95% of the region is outside this identified (Fig. 4). Note, however, that doing so will produce discontinuities at places where more than two boundaries meet. Furthermore, there is no guarantee that the confidence intervals for all types for a given location will sum to 1.0. That is, if I have a place that is located exactly at the 95% confidence level for Type A, this means that there is a 5% probability that this location is actually some other map type. Yet if I ask for the 95% confidence interval for Type B, it is possible that the same location will be located

Figure 4. Boolean map and 95% certainty map. Gray zones are less than 95% certain of being any given type.



within the area for Type B (Fig. 5a). Moreover, in the case of a Softwood/Hardwood boundary for a given Softwood polygon, the uncertainty would (presumably) be small meaning that the 95% confidence interval would be relatively close to the boundary on the thematic map. However, a Softwood/Mixed boundary for the same polygon would have much less certainty (presumably) meaning that there is a discontinuity even for the same polygon boundary (Fig. 5b). It is even possible that, if uncertainties are large enough for certain types, the very existence of a polygon in a given place is questionable (Fig. 5c).

There are other problems with this approach which are inherent in the way that the Uncertainty Library is developed. Of critical importance in this construction process is the size of the buffer zone selected for use. Not only does this affect which lines from the original interpretations will be bundled as representing the same "true" line, but it also affects the maximum uncertainty that will be found in the Uncertainty Library. For example, if the buffer zone selected is 20 m with three interpretations, then the maximum uncertainty will be 40 m — i.e., three lines spaced equidistant at 20 m which will be bundled together. Moreover, because the 20 m may not be applicable over the entire length of a line, one must decide subjectively when something is to be considered an outlier (Fig. 3). By definition, certain lines will be assessed as outliers even though this may be because they cause problems for the methodology and not because they are truly statistical outliers. The net result of all of these factors is that the values in the Uncertainty Library are likely to be underestimates of the actual spatial uncertainties associated with a given boundary line. Finally, this method must make the assumption of a Gaussian distribution across a line. While there is evidence to support this assumption for the synthetic images on which the method was originally developed, there has not been exhaustive testing of this under a wide variety of cartographic conditions.

3. Method 2: "Polygon core"-based Certainty Maps: Inside-Out

3.1 Data collection

Traditionally, photo-interpretation is conducted by identifying boundaries of homogeneous areas with the constraint that the boundaries form closed polygons over the entire map. Each polygon is then labeled with its appropriate cartographic appellation. In the proposed certainty-based photo-interpretation, it is only required that interpreters identify those features which are "100% sure." It is not necessary that these form closed polygons Theoretically, these features can be one of three elements (Fig. 6). First, one may have an actual, definite polygon. In forestry, such

Figure 5a. A situation in which the 95% certainty of one boundary is completely within the 95% certainty of another.

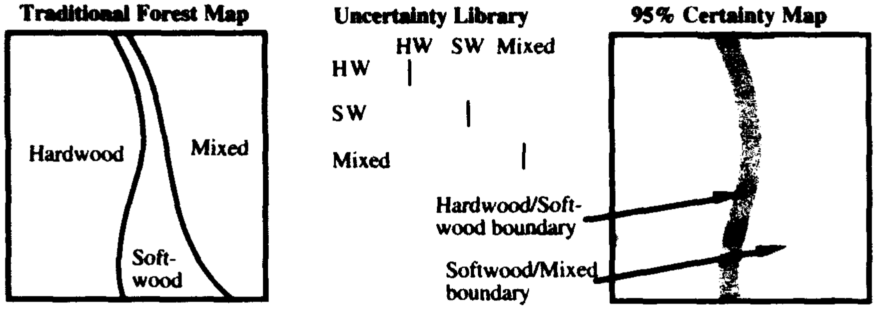


Figure 5b. A situation in which the boundary of a single feature has discontinuities.

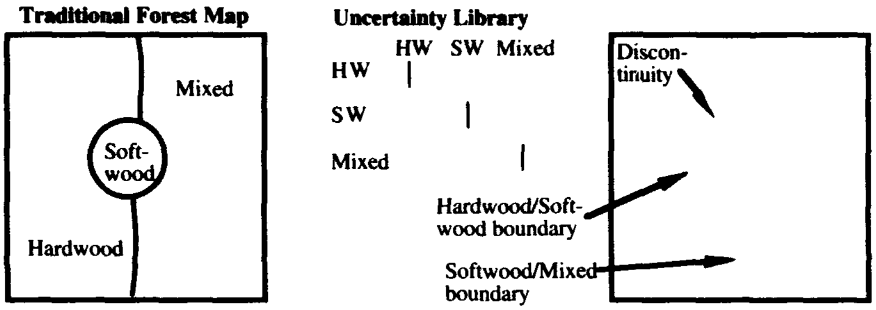
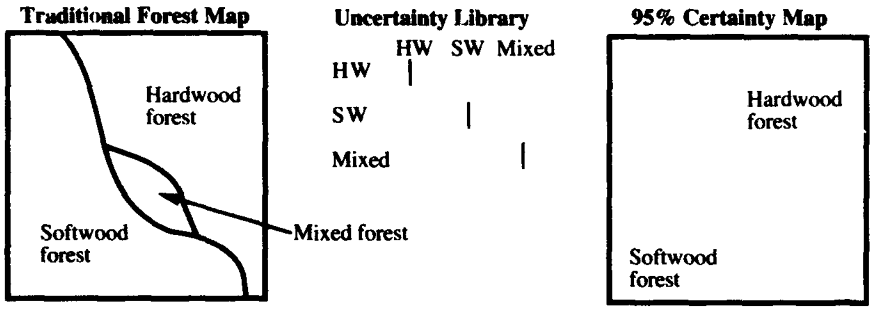
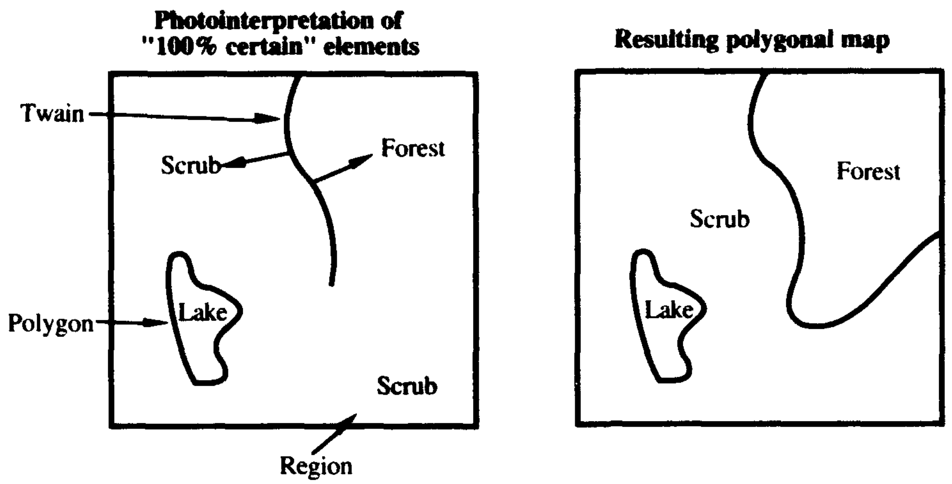


Figure 5c. A situation in which the magnitude of the uncertainty calls into question the existence of a polygon.



elements are most likely to be lakes, or clearcuts, or power line right-of-ways, etc. — i.e., distinct elements with definite boundaries that truly are polygons. Note that with such objects, only the interior of the polygon is known to be a given type; what is on the outside of the line delimiting the polygon is not necessarily identified. The second type of feature possible is a line for which the cartographic type on both sides of the line is labeled, but the inter-

Figure 6. Certainty-based photo-interpretation and polygonal map resulting from spatial interpolation and maximum likelihood classification.



preter is not obliged to form a closed polygon with the line. This type of feature is referred to as a "twain" herein. The third and final element possible is a region or point of a known cartographic type whose boundaries are not exact; instead it is the core of the polygon that is recognizable and identifiable with "100% certainty." Note that the resulting "map" has virtually no use for human interpretation because of humans being accustomed to closed polygon maps. It is critical that this information be treated or post-processed subsequently in order to render it useful for human interpretation.

3.2 Treatment and use of information

Treating this information requires a thorough understanding of the nature of the data. Effectively, a certainty-based map has a series of points labeled "100" (% certain of being a given type). In the case of a closed polygon on such a map, the interior and the bounding line are known to be, "without doubt," the cartographic type labeled. A twain is two sets of points side-by-side which is known to have one type "on the left" and a different type "on the right." Thus one has a set of "100% certain" points for one map type abutting a set of "100% certain" points for another map type arranged linearly. In effect, therefore, a twain acts as an impermeable membrane that prevents one type from "bleeding" into the other. Finally, in the case of a region — i.e., a polygon core — the boundaries of the region remain to be defined, but that core is a set of "100% certain" points for the map type labeled. Note that this way of looking at the data as a series of points labeled 100 also implies an equally valid inversion of the data. That is, if a set of points are "100% certain" to be Forest, then they are also "0% certain" to be Lake and/or Clearcut, and/or any other map type. Thus if one wants to produce a certainty map for Forest, one need only label all Forest points "100" and all others "0" and conduct a spatial interpolation.

If this is done for all map types individually, one obtains a certainty surface for each type which may be treated very similarly as the certainty map produced from multiple comparisons. For example, to produce a polygon map that identifies map type boundaries, one may do a maximum likelihood classification: assign each point to the class for which its certainty is the largest. Note that the boundaries so identified will effectively be the "50% line" between two classes, or the "33% line" among three classes, etc. One may also ask more specific questions of the certainty surfaces than just polygon boundaries. For example, one

may ask — as before — for the map showing 95% certainty for all types (Fig. 7). This request highlights a potential problem, however.

In performing the interpolation of uncertainty in the manner described, it must be assumed that the form of the distribution of certainty from one 100% certain element to another is linear — something that may not be true. The reason for this imposed linearity involves the interpolation method that is required. Normally when one interpolates spatially, one has a single variable or set of values — e.g., points of known elevation. However, in the present case one has a set of values for Lake, a set of values for Forest, etc. When the interpolation is conducted, not only is a certainty value needed for each type at every location, but the certainty values for a given location must sum to 100. Thus we interpolate in a seemingly typical fashion, but with an added constraint. The only method for doing this known to the author is area-stealing interpolation (Gold 1989) — a variant of natural neighbor interpolation. With this method, one effectively determines the certainty value for a given type by assessing geometrically the influence of all neighboring "100% certainty" points and their associated types. (For more detail see Lowell (1994).) Because one is not literally interpolating across a boundary, one cannot change the form of the distribution across the boundary.

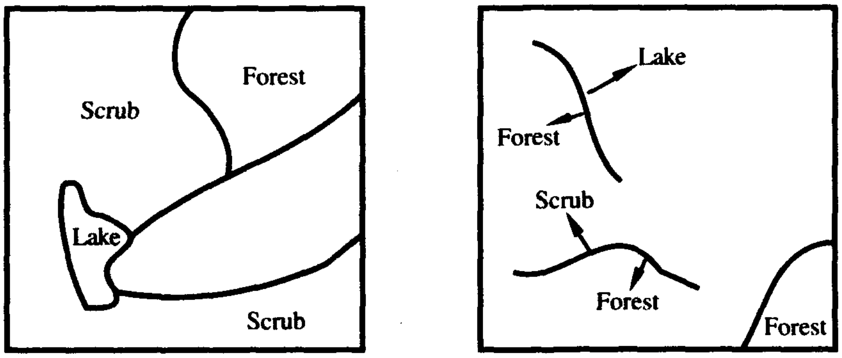
Another more subtle problem is in the nature of the data. In the example presented (Fig. 6), the data "made sense" and could be understood without the use of a computer. However, it is easy to imagine a situation in which the data do not "make sense" (Fig. 7). Nonetheless, because a non-intelligent algorithm will be applied to these data, a result/surface will be produced even though it may be nonsensical. Note that the problem is not with the algorithm employed; no interpolation algorithm is capable of understanding that certain data will not produce "meaningful" polygons. The problem is simply that the data make no sense relative to the way in which human beings interpret the world — something that is related to the desire to have homogeneous polygons — whereas the interpolation algorithm is certainly capable of using the data. In fact, ensuring a surface that "makes sense" is one of the reasons that conventional methods have been employed — there is always an internal check on the consistency of polygons. This is not the case with certainty-based interpretations which may cause problems for unwary users.

4. Synthesis and Conclusions

Two methods have been presented for producing certainty surfaces from data derived from subjective human inter-

Figure 7. Map showing areas which are "95% certain" (gray area is less than this for all types) and a certainty-based interpretation that would produce a nonsense polygonal map.

95% Certainty Map



pretation of a particular phenomenon. These are applicable to phenomena for which the production process is subject to considerable subjectivity.

One method re on the boundaries identified by an individual and a library of uncertainties. It also involves some subjectivity in that in producing the Uncertainty Library, the size of a buffer employed during one operation, as well as the determination of which interpretations or portions of interpretations are outliers, are subjective decisions that will vary from one individual to another. It also has the drawback that it makes an assumption about the form of the distribution of the error across a boundary once the Uncertainty Library is available. Furthermore, there are discontinuities in the nature of error at places at which three or more polygons join. And finally, it requires multiple interpretations of a phenomenon and/or the availability of a pre-existing Uncertainty Library.

The other method relies on data showing only those features which are 100% certain on a single interpretation and an interpolation algorithm. This method, though less subjective, also has inherent limitations and drawbacks. Simply the manner in which photographs must be interpreted is a drawback. Photo-interpreters are currently trained to identify closed polygons over an entire surface. Suddenly asking them to change their method of interpreting from "Identify closed polygons" to "Identify only those map elements that are 100% certain" is sure to cause a certain amount of discomfort and misunderstanding initially. This method also suffers from the impossibility of defining a particular frequency distribution for error as one moves from one "100% certain" element to another. Although relatively little work has been done on determining the true form of error distributions across map boundaries, it is certainly conceivable that this is not linear as must be supposed herein. Finally, this method can and will produce polygon surfaces from data even if the basic data are essentially nonsensical.

Despite these drawbacks for either method, the concept of developing certainty-based maps for interpreted phenomenon is sound. Regardless of the method of construction, such maps clearly provide more flexibility to a user than existing Boolean maps. However, it remains that, just as with Boolean maps, the certainty-based maps will only be as good as the assumptions and data used to construct them.

Bibliography

Aubert, E. (1995) Quantification of the spatial uncertainty inherent in photo-interpretations using a GIS for temporal monitoring of forest stands. Unpublished M.Sc. thesis, Université Laval (in French). 92 pp.

Edwards, G., Lowell, K. (1996) Modeling uncertainty in photo-interpreted boundaries. *Photogrammetric Engineering and Remote Sensing* 62:337-391.

Fisher, P. (1992) First experiments in viewshed uncertainty: simulating fuzzy viewsheds. *Photogrammetric Engineering and Remote Sensing* 58:345-352.

Gold, C. (1989) Surface interpolation, spatial adjacency, and GIS. In *Three-dimensional Applications in Geographical Information System*. Edited by J.R. Raper. Taylor and Francis, London. Pp. 21-35.

Goodchild, M., Guoqing, S., Shiren, Y. (1992) Development and test of an error model for categorical data. *International Journal of Geographical Information Systems* 6:87-104.

Hall, G.B., Wang, F., Subaryono, G. (1992) Comparison of Boolean and fuzzy classification methods in land suitability analysis by using geographical information systems. *Environmental Planning A* 24:497-516.

Lowell, K. (1994) A fuzzy surface cartographic representation for forestry based on Voronoi diagram area stealing. *Canadian Journal of Forest Research* 24:1970-1980.

Ministry of Natural Resources (1995) *Norme de stratification Écoforestière: troisième programme de connaissance de la ressource forestiere*. Ministère de Ressources naturelles du Québec, Service des inventaires forestiers. 116 pp.

The Web and Spatial Information Systems: Beyond Hypermedia?

D.J. Abel, R.G. Ackland and S. Hungerford
CSIRO Mathematical and Information Sciences,
GPO Box 664,
Canberra, ACT 2601,
Australia.
{dave.abel, ross.ackland, stuart.hungerford }@cmis.csiro.au

Keynote Speech *presented at the second annual conference of GeoComputation '97 & SIRC '97, University of Otago, New Zealand, 26-29 August 1997*

Abstract

The World Wide Web and the Internet have great potential in improving accessibility to spatial data and to spatial data processing services. We explore these themes by reference to two pilot systems. The ACT Pilot system assessed the feasibility of transferring data in vector form, to facilitate a wider range of modes of interaction than possible with delivery of data products as images. The SMART (Spatial Marketplace) Project is assessing the technical feasibility of Spatial Internet Marketplaces, in which applications are built from data and processing services offered by providers.

1. Introduction

The World Wide Web (the Web) has become an immensely valuable information resource. It is a compelling example of the synergy that can exist between new technology and the new applications enabled. The appeal of the Web stems from the combination of the near-global reach of the Internet, the ease of publication of information, and the simplicity of access by users to that information. Not surprisingly, there is high interest from Spatial Information Systems (SIS) researchers, practitioners and vendors in exploiting the Web. The primary exploitation to date has been to distribute, widely and conveniently, spatial data products. For example, a council might use an Intranet to distribute cadastral maps to its policy and operational units, without having to equip them with specialist software. At another level, map-like visualisations are compact and informative reports which are likely to be used by a wide variety of Web information providers.

In this paper, we consider two themes in the use of the Internet by the SIS community. The first is extended use of the Web for distribution of spatial data, particularly to allow the richer set of interactions with spatial data expected in conventional GIS but not readily possible with the standard Web tools. The second is access, using the Internet, to specialist spatial data processing services.

There are already many impressive Web sites offering spatial data products. These first-wave combinations of SIS technology and the Web deliver spatial data as hypermedia documents. This approach is relatively simple technically. At the server, an image (in GIF format) is generated, either by generating a vector map and converting to an image or by extraction from a database of maps held as georeferenced images. At the browser, the image can be viewed using the standard tools for displaying images. Specialist SIS tools are needed only on the server. This strategy is effective for the many applications requiring only standardised products. It is less effective when there is a need for interaction with the spatial data or for extensively-customised displays. For example, it is not readily possible to turn layers on and off, or to click on a feature to obtain further information. These operations can of course be accommodated by using clickable images to launch a new request to the server for a new document. It is attractive, then, to consider more sophisticated ap-

proaches which offer a richer set of interactions and possibly lower data communications costs. We explore some approaches to this issue in Section 2 through analysis of the design and performance of a system for Web distribution of cadastral and other data, particularly in linking property maps with such aspatial data as ownership, sales and valuation information.

It is also of interest to consider how the Internet might be used to allow an application to draw on remote processing services. For example, a geographically-dispersed organisation might prefer to mount specialised software on a high-performance computer for use from all its sites, avoiding the costs of installation and maintenance at all sites. In other cases, the motivation is to access specialist software and hardware which is used too infrequently to justify purchase. In Earth Observation, for example, many operations require extensive processing and some applications call for rapid generation of products. An analyst tasked with estimating wheat production might prefer to use High-Performance Computing facilities at a remote site for geometric correction and enhancement rather than more limited local facilities. The question, then, is how the Internet might be used to enable access to such services. In Section 3, we present the concept of a Spatial Internet Marketplace, essentially an infrastructure for the publication of processing and data services. An architectural model is offered and is demonstrated by a simple application for itinerary planning. In Section 4, we consider the more general implications of the new or modified application of Spatial Information Systems technology enabled by the Web and the Internet.

Distribution of Spatial Data

2.1 System Architectures

The Web is essentially a client-server environment specifically designed as a distributed hypermedia information system. The standards for interfaces between clients and servers in the Web are relatively light, and permit a large number of approaches. A useful first-level categorisation of the approaches is in terms of the assignment of functions between the client and the server. The extremes are then

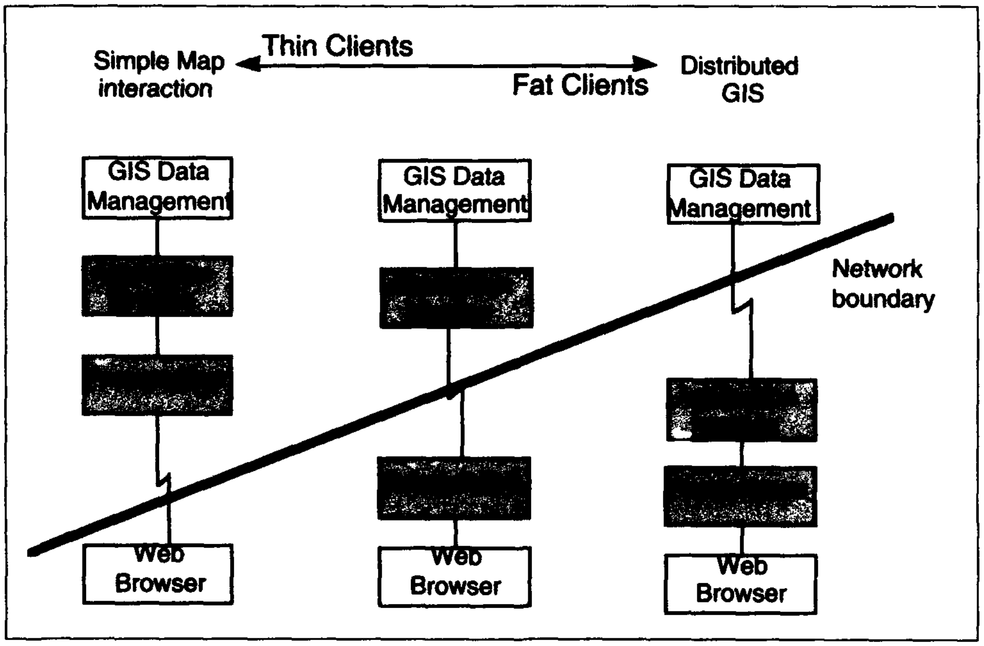


Figure 1 Assignment of Functions in Distributed Systems

the 'thin-client, fat-server' and the 'fat-client, thin-server', whereby 'thin' implies a minimal amount of installed software and 'fat' implies a relatively large amount of installed software needed to perform the task Figure 1 illustrates some possible assignments for spatial applications.

The simplest aproach is the thin-client. The client system is the Web browser (Netscape or similar), so that there are no unusual demands on the hardware or software installed by the user. To conform with the types of hypermedia documents accepted by the browser, the spatial information is delivered as a GIF image. Where the data product has fixed content (i.e. a pre-specified collection of layers) and requests need specify only the geographic region of interest, the image for the whole region can be pre-materialised and stored in a database or a native file system (e.g. Lamb, 1994) as a set of tiles. Retrieval for a specified region then extracts the tiles intersecting the region, assembling them into an image, and clipping the image. Where the content varies, the image can be generated by extracting from a database the features of interest and generating the image on-the-fly. Some representative sites following this approach are http://pubweb.parc.xerox.com/map, http://www-nmd.usgs.gov/index.html, and http://www.erin.gov.au/database/db.html.

Extending the operations provided by the browser (and so providing a richer set of interactions with the data once delivered to the client) clearly requires a 'thicker' client. The two more common approaches used to enhance the capabilities of the browser are plug-ins and applets. A plug-in is essentially an application, able to be launched by the browser, which accepts data of a known type and representation and provides the specialist viewing and data manipulation operations for that type and representation. The plug-in is installed prior to its use. An applet, on the other hand, is a program delivered by the server along with the data. Applets are coded in the Java language and are executed interpretively (or compiled "Just In Time" after they are downloaded). Applets also have the advantage of being machine independant compared with plugins that are built for specific platforms only. Both approaches allow the development of tools for a wide range of operations on spatial data. Importantly, they allow use in Web environments of data types beyond the mainstream hypermedia types. For spatial data, this allows use of data in vector representations. This, of course, requires definition of file transfer formats for the spatial data. For some representative examples of plug-ins, see http://www.softsource.com/ and http://www.mapguide.com/. For applets, see http://maps.purple.org/map/index.html and http://www.neosoft.com/~forge/java/Cartog/Cartog.html.

An applications configuration now possible is that the server is quite thin and the applications functions and data manipulation is handed over to the client. If the manipulation of the data is performed by the client, the server can be implemented using a Commercial-off-the-Shelf spatial database engine with a wrapper to provide HTTP connections, to interpret requests from clients, and to load the retrieved data into the required file format. This is at the expense of the memory needed for the applets or plug-ins at the client. Where an applet is used, interpretative execution can also have a performance penalty, especially when manipulating or drawing large volumes of data.

To determine the feasibility and effectiveness of a fuller set of facilities for viewing and manipulating spatial data in Web environments, CSIRO Mathematical and Information Sciences (CMIS) has conducted two pilot projects. Both sought to explore the technical issues in implementation and to test the value of solutions by adopting 'real world' problems.

The ACT Pilot

The ACT Pilot Project, undertaken collaboratively with the ACT (Australian Capital Territory) Land Information Centre, assessed Internet access to the core government cadastral and related databases. These data sets are important as they are referred in a great many administrative, planning and operational processes. Internet access is also potentially attractive as an alternative to delivery of the information to industry and the community in hard-copy form. Technically, the project assessed the feasibility of delivering the data in vector form, providing zoom and

Query Type	N Objects	N Vertices	DWF File Size (Kbyte)	GIF File Size (Kbyte)
Block	36	3112	21.1	12.8
Street	377	8004	75.6	27
Area	1338	14455	91.1	32.2
Suburb	6810	52492	332.7	36

Table 1 Content of the Four Test Queries

pan, switching layers on and off, and clicking on features to provide further information.

The databases included the Digital Cadastral Database (DCDB) for the ACT; ownership, valuation and sales history for land parcels; navigational features such as roads and watercourses; and some other sets such as landmarks and footprints of certain classes of buildings. Summary characteristics of the major databases are shown as Table 1. The databases were installed from the ACT's ACTMAP and PALMS databases in Spatial Data Manager (SDM) (Abel, 1989) and ORACLE. The major data sets were roads (37398 objects, average of 8.3 vertices), contours (580063, 30.7), building footprints (37253, 8.9), land parcels (105813, 6.3) and annotations (321617, 4.0).

The broad design of the system is shown as Figure 2. The primary spatial query functions provided by the server were retrieval of objects within a region a (block) specified as a 20-metre square centred on a block identified by its street address, within a similar region (a block) of 200 metres sidelength, within the minimum bounding rectangle of a street, and a suburb (identified by name). Data was delivered to the client in the Autodesk Drawing Web DWF Format (see http://www.autodesk.com/products). The basic drawing functions (including zoom and pan) were performed by an applet implemented in Java. Additional GUI functions specific to the application were implemented using the Netscape Internet Foundation Classes (IFC) class library. Point-and-click queries were handled by tagging each feature with its identifier, and passing back to the server a query on the database using that identifier. Figure 3 shows a sample screen.

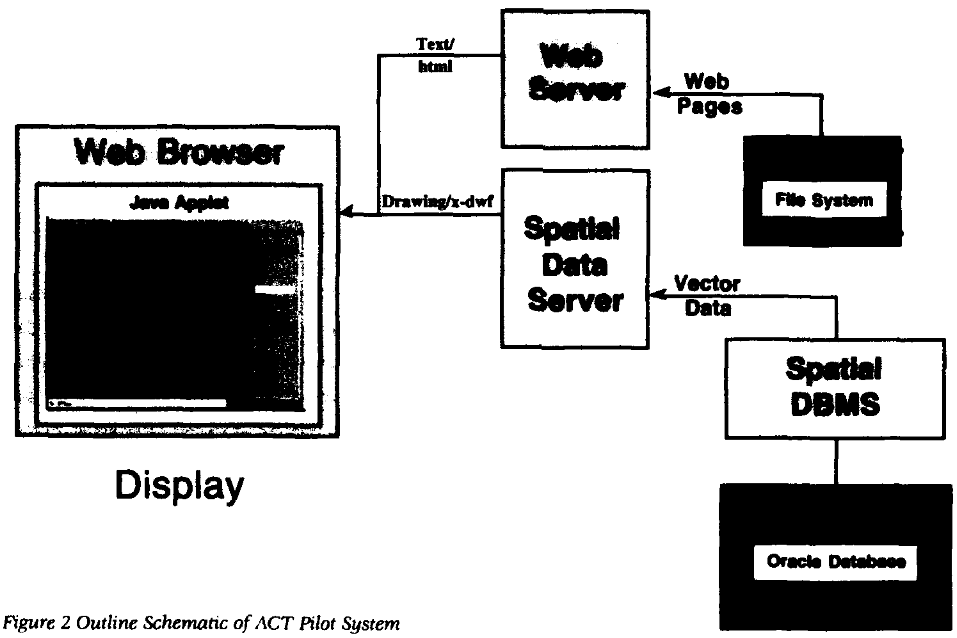


Figure 2 Outline Schematic of ACT Pilot System

Query Type	Data Extract	DWF Load	Network Transmit	Applet Draw
Block	0.62	1.00	0.03	0.13
Street	3.43	3.16	0.13	0.61
Area	7.39	6.91	0.14	1
Suburb	8.19	30.69	0.99	6.25

Table 2 Breakdown of Costs

The investigation confirmed that the required operations could be implemented and that the operations in the browser (once the DWF file was received) were genuinely interactive with performance very similar to that for a conventional GIS, after some care in the Java coding to optimise the graphics display performance The hardware configuration was a SUN 5/700 as the server and a 200 Mhz Pentium PC under Windows NT 4.0 as the client with Netscape 3.0. The network was Ethernet rated at 10 Mbit/s.

We report some performance data in terms of four representative queries. Summary characteristics, the size of DWF file and the size of the corresponding GIF files to generate the same displays are shown in Table 1. The breakdown in the elapsed times for the steps in generating the vector display, with an Ethernet LAN of 10 Mb/s, are shown in Table 2. (Note that these figures indicate that the LAN was operating at an effective 2.7 Mb/s to 5 Mb/s.)

For these queries, the DWF format has a distinct size disadvantage which worsens with higher data content. The DWF file size increases, almost linearly, with the number of vertices while the GIF file is relatively insensitive to content. For an Intranet environment with a LAN, this size disadvantage is relatively unimportant: it translates to a

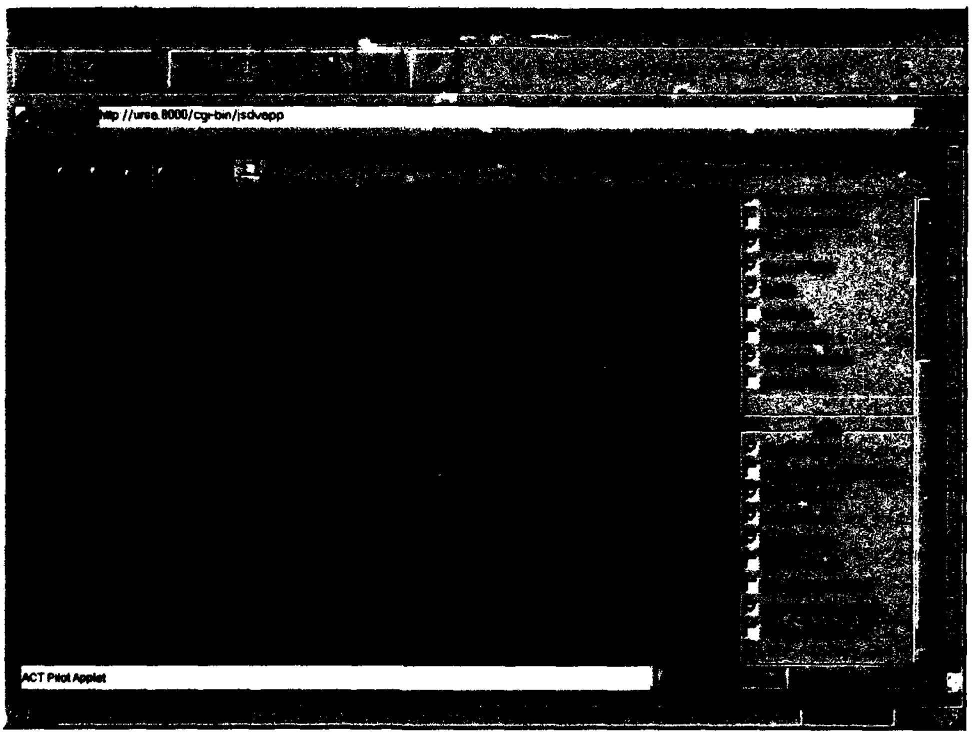


Figure 3 Sample Screen from ACT Pilot

network transmission cost penalty of 0.16 seconds to 0.97 seconds. For an Internet dialup connection rated at 28.8 Kb/s, the cost differences are higher by a factor of 100, and the penalty in transmission costs becomes significant. Some conditions specific to the data used suggest caution in drawing general conclusions from these comparisons. The sample screen shows a very large number of circular arcs for parcel boundaries and street centrelines and road casements. The prototype system rendered these as sets of small line segments and so increased the sizes of the DWF files. Additionally, the displays include a large amount of unfilled background, which favours GIF images.

The breakdown of the elapsed time for the basic operations, however, shows that total costs are dominated by the server -resident operations, for LAN/WAN environments. The draw times, for the smaller displays, are genuinely interactive. The response times for the street and area queries appear within the times needed for operational use. The suburb query, for the prototype implementation, is marginal, even for use with LAN/WAN networks. It is probable that further development would cut the costs significantly. In general, the DWF file emerges as the design choice most deserving reconsideration. It appears quite likely that investigation of transmission formats designed specifically for spatial Internet applications will find formats which are smaller and less expensive to encode.

3. *Spatial Internet Marketplaces*

Clearly a client could accept data in a DWF form (for example) and use it freely as input for local processing. A supermarket chain, for example, might download demographic data for the regions around its outlets, and use that data together with its private sales data for outlets to compare the performances of the outlets. That is, the Web can readily support the publication of data services in addition to data product services. A further extension is to consider the publication of processing services. Here a service provider would accept from a client an input data set and specifications of the processing to be performed, perform the processing at the provider's site using the provider's software, and return the outputs to the customer.

This follows the concept of electronic marketplaces advanced for decision technologies (e.g. Bhargava et al, 1995; Guenther et al, 1996). The primary motivation was that accessibility to software systems was limiting the adoption of the technology by business and industry. The solution envisaged was that a number of service providers would publish, on a fee-for-service basis, analytical and optimisation services. Customers would then identify suitable services for a task and choose the best service for their purposes. A complex investigation might involve use of services from several services. The overall operation of the marketplace- the entry of providers, the sophistication of services, the charges levied, and so on- would be driven by market forces. Broadly a marketplace would consist of a number of service providers, a number of customers, and an *infrastructure of registries* of available services and a certain set of standards.

More recently it has been proposed that Spatial Internet Marketplaces (SIM's) are potentially attractive to the SIS community (Abel 1997, Abel et al 1997). The concept remains close to that of the electronic marketplace. In addition to processing service providers offering specific transformation, data fusion, analytical, simulation and optimisation services, there would also be data providers. A customer could then buy data from one provider and route it to other providers for processing, receiving only the products required. To assist in finding desired data and processing services and in developing a sequence (a plan) of service invocations, facilitator services would be available.

3.2 The SMART Architectural Model for a SIM

The SMART (Spatial Marketplace) Project at CMIS is aimed at establishing an architectural model for a SIM infrastructure and at exploring the implementation of the model through development of a series of pilot systems. The project is particularly aimed at identifying forms of infrastructures which provide a balance between ease of use for customers (by controlling the degree of variation be-

tween providers) and the costs of participation as a provider. We see the definition of a lightweight set of standards as crucial.

A significant design decision has been to restrict the interaction between a customer and a provider to be coarse-grained (equivalently, stateless). That is, the customer makes a request to a provider, notionally as a single message, who then delivers the result to the customer, again notionally as a single message. Any further requests by the customer to the provider is treated, fully, as separate requests. This is in contrast to fine-grained interaction, where the customer establishes a connection (or session) with a provider, passes requests, receives data from a request in possibly several messages, and closes the connection when no more requests are anticipated.

The SMART architectural model has two types of services. A **query service** provides retrieval from a store of pre-materialised data. A query service would typically be implemented using a database system. It is defined in terms of its schema (the descriptions of the sets of entities and their attributes) and the functions and operators able to be included in predicates. A **function service** derives data from a data set supplied by the customer, possibly by reference to further data and information held by the provider. For example, a function service to convert Australian dollars to New Zealand dollars might refer to a conversion rate for differing amounts to be converted held by the service. A function is described by a schema which gives the permissible combinations of input data, the outputs which can be generated, and the domain of applicability of the function service (i.e. the presented problems which it can reliably solve).

A service (either a query or a function service) is invoked by passing to it a command expressed in the SMART Request Specification Language (the RSL). The major motivations for yet another language are applicability to both query and function services and the avoidance of the complexity of languages such as OQL and SQL which increase the costs of participating for providers. An RSL statement has two parts: the constraint specification and the target list specification. The constraint specification essentially declares the set of objects of interest, and the target list the attributes of those objects to be reported to the customer. Examples of RSL statements on a query service and a function service (respectively) are:

```
suburb.name = 'Dickson', motel.location
    withinsuburb.location,
    motel.rating = '***' # motel.adress, motel.name;

stack.site = (56, 12000, 13000), discharge = 13.2,
        time =(1...100) # stack.plume;
```

The SMART infrastructure includes a special query service, the **Registry**, and two special types of function services, **Planners** and **Executors**. There is a single Registry, which has a role closely similar to that of an Object Management Group Trader. It stores the external schemas of all other services and of itself (describing the precise data and operations available to customers), a type registry of data types and operations, a thesaurus of terms available to describe entities, attributes and types, and a glossary documenting the terms. The information held in the Registry can be queried by invoking the Registry's own query service. A mandatory requirement for participation in the SMART marketplace is that providers register their services by providing (and having accepted) descriptions of their services.

An Executor accepts a **Plan** (a sequenced set of requests on services, together with some conditional and other statements) and executes it on behalf of the customer. A Plan, in its simplest form, is a Java program and an Executor a Java interpreter. A Plan can be developed manually by an applications developer and stored for repeated use. A Planner accepts a statement of data required by a customer and generates automatically a Plan which will materialise the data.

3.2 The ACT-TAP Pilot

As a first test of some of the core elements of the SMART infrastructure design, an experimental system, ACT-TAP (Australian Capital Territory - Tourism Advisory Project), has been built. This is not yet an implementation of all

components of the infrastructure, in even the most basic forms. Rather the emphasis in ACT-TAP has been on assessing the difficulty establishing services and of invoking those services. More generally, ACT-TAP is intended as a framework for further research on the more complex aspects of the infrastructure.

The marketplace established (Figure 4) has two query services and three function services:

- the KDB (Kerry's Database) service is a database of features and facilities in the ACT of interest to tourists, such as restaurants, hotels and attractions such as the Australian War Memorial and the Houses of Parliament. Each facility has a street address. Some have URL's for web sites with further information. The KDB is implemented as an ORACLE database;
- the ACTSpatial query service includes the databases of the ACT Pilot Project. This includes the road segments of the ACT, tagged by identifiers but not topologically connected;
- the RoadNet is a function service which evaluates the shortest path using the road networks of the ACT. It takes an origin and a destination point (represented by their Australian Map Grid coordinates) and returns the shortest path between them (as the sequenced list of identifiers for road segments) and the estimated transit time. It is based on a network representation of the ACT road network, in which each node is the junction of two roads and is tagged with its Australian Map Grid coordinates and each arc is the connecting road segment;
- the Scheduler service solves routing-and-scheduling problem. It accepts a list of sites to be visited, a matrix of the travel times between sites, the earliest and latest times to arrive at each, and the time to be spent at the site. It uses a tree search to evaluate the sequence of sites which minimises the total travel time;
- the Itinerary service solves a routing-and-scheduling problem. In this case, the input is a list of names of sites and features, and the earliest and latest times. It returns the itinerary which minimises the total transit time.

The ACT-TAP application notionally assists a prospective visitor to the ACT to plan a visit by building a list of sites

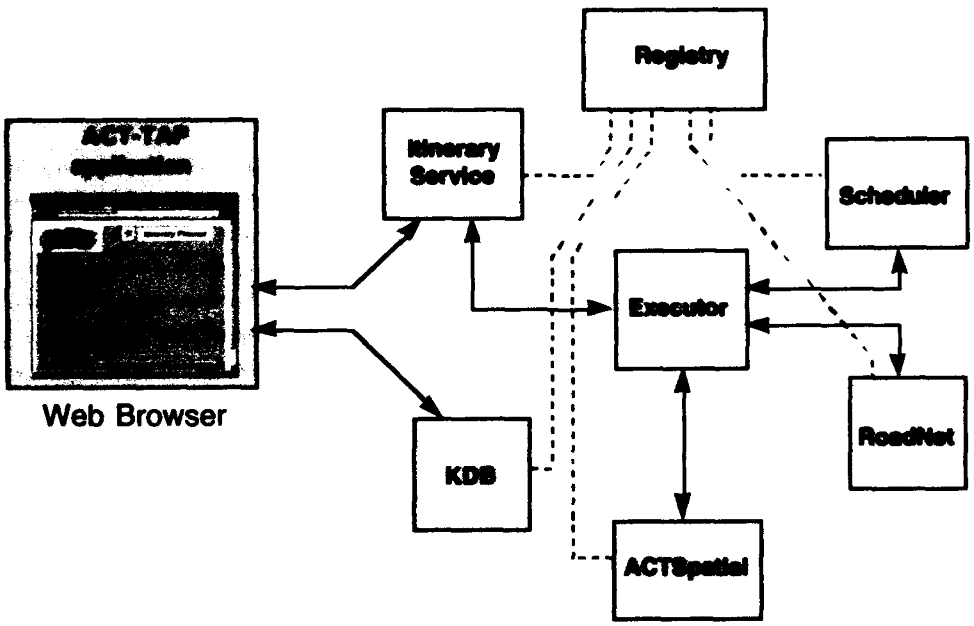


Figure 4 The ACT-TAP System

to visit and selected hotels, restaurants, and so on, and then suggesting an itinerary which minimises the time travelling between them. ACT-TAP application itself is implemented as an applet executed within a browser. The applet allows the user to interrogate the KDB and enter, on a form, the selected list of sites and facilities. The user can also access, through the browser, the Web sites referenced in the KDB information. When the list is complete, the applet invokes the Intinerary service:

- The Itinerary service simply fetches a stored plan, and passes it to the Executor. The plan simply invokes the query and function services, with some reassembly of data to suit the external schemas of the services;
- the KDB service is invoked to fetch the street addresses of the sites and facilities;
- the ACTSpatial Service is invoked to determine internal points for the land parcels corresponding to the street addresses;
- the RoadNet service is invoked to determine the distance matrix;
- the Scheduler is invoked to evaluate the minimum-distance itinerary.

The ACT-TAP application notionally assists a prospective visitor to the ACT to plan a visit by building a list of sites
to visit and selected hotels, restaurants, and so on, and then suggesting an itinerary which minimises the time travelling between them. ACT-TAP application itself is implemented as an applet executed within a browser. The applet allows the user to interrogate the KDB and enter, on a form, the selected list of sites and facilities. The user can also access, through the browser, the Web sites referenced in the KDB information. When the list is complete, the applet invokes the Intinerary service:

- The Itinerary service simply fetches a stored plan, and passes it to the Executor. The plan simply invokes the query and function services, with some reassembly of data to suit the external schemas of the services;
- the KDB service is invoked to fetch the street addresses of the sites and facilities;
- the ACTSpatial Service is invoked to determine internal points for the land parcels corresponding to the street addresses;
- the RoadNet service is invoked to determine the distance matrix;
- the Scheduler is invoked to evaluate the minimum-distance itinerary.

4 Relationship to Other Work

The Spatial Internet Marketplace is clearly closely related both to topics in Spatial Information Systems technology and applications. The functions of the infrastructure are similar to those of the core elements of workflow s tems and of multidatabase systems, and can also b lated to the more general components for distributed architectures under the CORBA model (for example). A key difference is the absence in the marketplace of a global schema. While the Registry contains the collection of external schemas, these are not integrated.

There has been extensive work on catalogues of spatial data collections which are network accessible. The New Zealand GUILD system is a representative example. In many ways, the marketplace's Registry is similar to a metadata catalogue, and we envisage that the metadata catalogues will be an important starting point for establishing marketplaces. There has been less work on using catalogues in accessing the data, although Pascoe (1996) describes a system including software agents to assist formulation and execution of a search and to acquire and transform data. The architectural model for a marketplace is conceptually very similar, with the major differences the consideration of both query and function services and the provisions for automated planning of query execution.

5.- Conclusions

The Web has already had a major impact on Spatial Information Systems through widening the accessibility of spatial data. The broadened market base for SIS practitioners and researchers will, almost certainly, be influential as a driver for research and development. There remain some applications which are handled only with great difficulty with the current Commercial-off-the-Shelf (COTS) tools.

A challenge for researchers and practitioners will be to determine where these applications can be better addressed by innovative use of COTS technology and where the technology is desirably extended. The ACT Pilot study suggests that it is indeed feasible to extend the technology while remaining compatible with the Internet and Web environments.

We do not argue, however, that these new approaches are in some way universally-superior replacements for the current tools. Rather they are offer new options in balancing modes of end-user interaction, transmission costs and server characteristics. They are then most accurately viewed as extending the range of choices open to an applications developer in matching choices of tools to the users' requirements, case-by-case.

The Spatial Internet Marketplace is clearly inspired by the Web model and the research reported here is essentially aimed at testing whether spatial data and spatial data processing can similarly be made widely (if not freely) accessible. The early results from the SMART Project suggest that the goal is technically possible. It remains to be seen if a critical mass of customers and providers could be achieved.

Acknowledgments

The ACT Pilot and SMART Projects have been pursued by a team of which the authors are members. This paper draws on the contributions of our colleagues in the Spatial Information Systems Project of CSIRO Mathematical and Information Services, and we gratefully acknowledge the many contributions of Volker Gaede, Dean Jackson, Rob Power, Gerritt Riessen and Xiaofang Zhou. We are also indebted to colleagues from industry and government who have alerted us to challenging research issues and who have supported our work.

References

Abel, D.J. (1997), Spatial Internet Marketplaces: a Grand Challenge, Proceedings of 5th Intl. Symp. Large Spatial Databases, to appear.

Abel, D.J., Gaede, V.J., Taylor, K.L. and Zhou, X. (1997), SMART: Towards Spatial Internet Marketplaces, Technical Report SIS-97-04, CSIRO Mathematical and Information Sciences, Canberra, June 1997.

Bhargava, H.K., King, A.S. and McQuay, D.S. (1995), DecisionNet: an Architecture for Modelling and Decision Support over the World Wide Web. In Tung X. Bui (Ed.), Proceedings of the Third International Society for Decision Support Systems Conference, Vol II, pages 541-550, Hong Kong, 1995, International Society for DSS.

Guenther, O., Muller, R. Schmidt, P., Bhargava, H. and Krishnan, R. (1996), MMM: a WWW-Based Method Management System for Using Software Modules Remotely, Discussion paper 32-1996, Sonderforschungsbereich 373, Humboldt-Universitaet zu Berlin, 1996.

Pascoe, R. (1996), Data Sharing using the X.500 Directory, Proceedings Conference on Geocomputation, pp 689-698.

AGPCS - An Automatic GSM-based Positioning and Communication System

Zoran Salcic
The Auckland University,
Department of Electrical and Electronic Engineering,
20 Symonds St., Auckland, New Zealand
E-mail: z.salcic@auckland.ac.nz

Presented at the second annual conference of GeoComputation '97 & SIRC '97, University of Otago, New Zealand, 26-29 August 1997

Abstract

A novel approach to automatic positioning and communication is presented in this paper. The approach is using single GSM (Global System for Mobile Communication) technology to achieve both positioning of the mobile station and communication with the other parties in a system containing a number of mobile stations. The paper deals with the overall system architecture, and briefly describes the aspects of positioning, communication and application of the system, without describing the low-level details applied in system operation. It also describes implementation of two experimental versions of the mobile stations, self-positioning and remote positioning, that integrate positioning information in the platform ready to use for different purposes. The approach is applicable to GSM and its other more recent derivatives.

1. Introduction

Automatic positioning (estimating location) and data communication is of great importance for many areas and applications, such as automatic vehicle locating and tracking, remote equipment and property locating and monitoring, boat, yachts, and cargo tracking and monitoring, remote patient tracking and monitoring, various types of dispatching and distribution systems, etc. Integration of positioning and communication into a single system is a goal not achieved in the contemporary positioning systems. Hence, it is the aim of our AGPCS (Automatic GSM-based Positioning and Communication System) project to integrate both positioning and communication based on the single GSM technology.

1.1 Global Positioning System

The most accurate positioning today is achieved using satellite-based global positioning system (GPS) (Kapplan 1996). However, the GPS has two important disadvantages. First, the information on position usually has to be transmitted to some other party requiring that the mobile station provides data communication facilities. Most often, it is provided using some sort of radio system based data transfers such as radio modems which communicate with specialised radio system infrastructure (such as trunked radio), or using an existing public cellular system. In the former case, this introduces the problem of radio coverage, and requires investment into radio transmission systems. Further costs are incurred by data communication devices. In the latter case, the positioning and communication facilities are at present implemented by two separate devices. The second disadvantage of GPS is that it is usable only for the case of "clear sky", which makes it hardly usable in urban areas, mountainous terrain, and in closed/covered space.

1.2 Radio Signal Propagation Models

Other methods for mobile station positioning are based on radio signal propagation such as signposts, dead reckoning, circular or hyperbolic trilateration systems, etc. Many methods and systems have been proposed based on radio signal strength measurement (Figel 1969, Ott 1977, Hata

1980) of a mobile object's transmitter by a set of base stations. Recently, adaptive schemes based on the use of cellular systems and on fuzzy logic (Song 1994), hidden Markov models, and pattern recognition methods (Kennemann 1994) have been used to estimate the position of mobiles. The most recent one (Hellebrandt 1997) is based on a multidimensional scaling technique. A mobile's position is determined in a such way that the measured signal strength of a certain base station in the GSM system is best fitted to the known average signal strength at this point. The performance of the method was tested by simulation for different simulated scenarios (Hellebrandt 1997).

1.3 Personal Communication Systems

The personal communications industry is one of the fastest growing industries in this decade (Feher 1995). The cellular market, as a part of this industry, is growing at the rate of almost 50 percent a year. The market offers a huge opportunity for many industries, including network service providers, software and hardware developers, and those who will upgrade the services offered by the basic network service providers. GSM, or other technologies derived from it, has become one of the prevailing cellular technologies worldwide.

1.4 GSM

GSM (Scourias 1995, Redl 1995) initially handled basic voice services and some emergency calling features, but has already added improvements in subscriber identity module (SIM) cards, which contain a microchip with the information on the caller. From the user point of view, the obvious difference between GSM and other cellular technologies is that GSM cellular phones operate only digitally, enabling both voice and data to be transferred directly digitally, without using modems, providing the backbone of the mobile communication network.

A variety of data services are offered in GSM. GSM users can send and receive data, at rates up to 9600 baud, to users on POTS, ISDN, Packet Switched PDN, and Circuit Switched PDN using variety of access methods and protocols. Other data services include G3 facsimile, and Short Message Services (SMS) which is a bi-directional service for short alphanumeric (up to 160 bytes) messages. Messages are transported in a store-and-forward fashion. For point-to-point SMS, a message can be sent to another subscriber, and an acknowledgment of receipt is provided to the sender. SMS can be used in a cell-broadcast mode, for sending messages such as updates of different sorts. Messages can also be stored in the SIM card for later retrieval. The SMS service provides a basic means to transfer data used to estimate position or coordinates of the mobile station.

Besides voice and data services, GSM system provides data that might be used for radio signal strength measurements and positioning. The GSM mobile station receives each 0.48 seconds the downlink signal levels from the serving and up to six neighbouring base stations in a discrete scale. The GSM mobile station applies a complex signal processing algorithms to determine the signal strengths. This information is a part of GSM system and is used in our system to estimate position of the mobile station.

1.5 New Approach - the AGPCS

By using, combining and integrating two inherent features of the GSM system (measurements of radio signal levels and ability to communicate directly digitally), we propose a novel Automatic GSM-based Positioning and Communication System (AGPCS) technology. The AGPCS is a real-time system built on top of the GSM system, and can be considered as an application layer to standard GSM. The first working versions of systems using AGPCS technology have been developed and tested. The AGPCS technology can be used for various applications, including control, as it may be easily incorporated into the standard hardware/software environments or used in the embedded form.

Our first goal was to obtain technology to estimate mobile station position with the accuracy that can be considered sufficient for a number of applications. The current model provides accuracy, which is almost always below

270m, and usually around 200m. Integration of position estimation and communication between mobile objects, or between mobile and stationary objects, has become feasible now. An integrated positioning and communication system, which can incorporate control features as well, is being obtained and used in a number of pilot applications. The AGPCS, including a brief overview of key technologies that take part in its implementation, is described in this paper.

2. The AGPCS Framework

In this section the AGPCS framework is described. First, we introduce some features of the GSM that are relevant for AGPCS. Then, the architecture and main features of the AGPCS are described.

2.1 Brief Overview of Relevant GSM Features

A GSM network is composed of several functional entities. It is illustrated in Figure 1., which shows the layout of a generic GSM network. The network is divided into three major parts:

1. The GSM mobile station subsystem.
2. The base station subsystem that controls the radio link with the mobile station
3. The network subsystem performs the switching of users and mobility management in the mobile services switching center.

The mobile station and base station subsystem communicate across the Um, or radio link, interface. The base station subsystem communicates with the network subsystem across the A interface. The International Telecommunication Union allocated the bands 890-915 MHz for the uplink (mobile station to base station) and 935-960 MHz for downlink (base station to mobile station). GSM is using a combination of Time- and Frequency-Division Multiple Access (TDMA/FDMA) method. At the 900 MHz range, radio waves bounce off everything - buildings, hills, cars, airplanes, etc. Thus, many reflected signals, each with a different phase, can reach an antenna. Equalization is used to extract the desired signal from the unwanted reflections. It works by finding out how a known transmitted signal is modified by multipath fading, and constructing an inverse filter to extract the rest of the desired signal. To minimize co-channel interference and to conserve power, both the mobile and the base transceiver station operate at the lowest power level that will maintain an acceptable signal

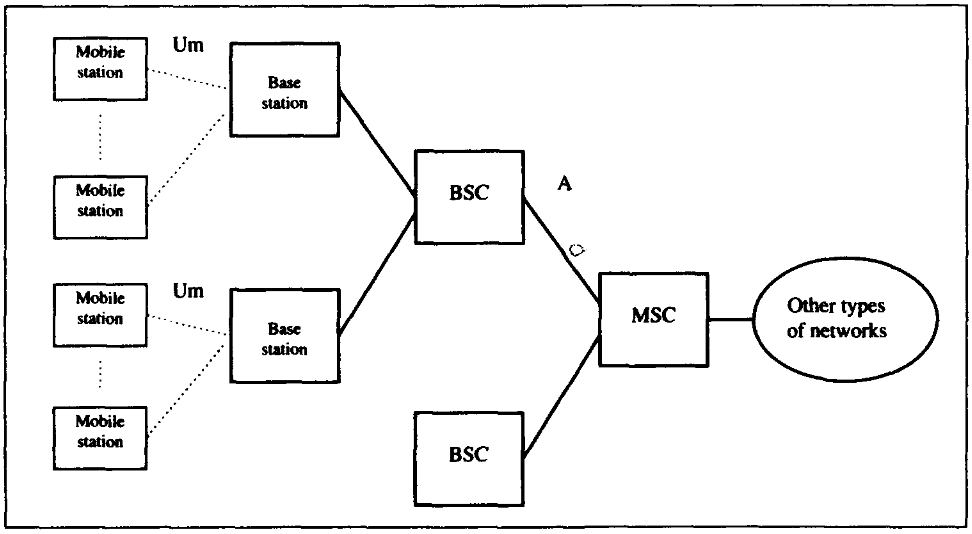


Figure 1. General architecture of a GSM network

quality. The mobile station measures the signal strength or signal quality (based on the Bit Error Ratio), and passes the information to the base station controller (BSC), which decides if and when the power level should be changed. Besides ensuring the transmission of voice and data of a given quality over radio link, the functions of mobile cellular network are the implementation of a handover mechanism, registration, authentication, call routing and location updating functions.

The signaling protocol in GSM is structured into three layers. Layer 1 is the physical layer, which uses the channel structures of GSM. Layer 2 is the data link layer. Across the Um interface, the data link layer is modified version of the LAPD protocol used in ISDN, called LAPDm. Across the A interface, the Message Transfer Part layer 2 of Signalling System Number 7 is used. Layer 3 of the GSM signaling protocol is itself divided into three sublayers:

- Radio Resource Management which controls the setup, maintenance, and termination of radio and fixed channels, including handovers. The management of radio features such as power control is performed in this sublayer.
- Mobility Management which manages the location updating and registration procedures, as well as security and authentication
- Connection Management which handles general call control, and manages supplementary services and short message service.

Obviously, from the AGPCS point of view, the most important is Layer 3, at which the AGPCS technology is hooked-up to GSM.

2.2 AGPCS Architecture and Operation

The AGPCS represents an application technology built on the top of standard GSM. It performs the positioning of the mobile station in the coverage area of the GSM network. The logical structure illustration of an AGPCS system is given in Figure 2.

The AGPCS mobile station consists of the GSM mobile station (actually handset) and a mobile computer connected

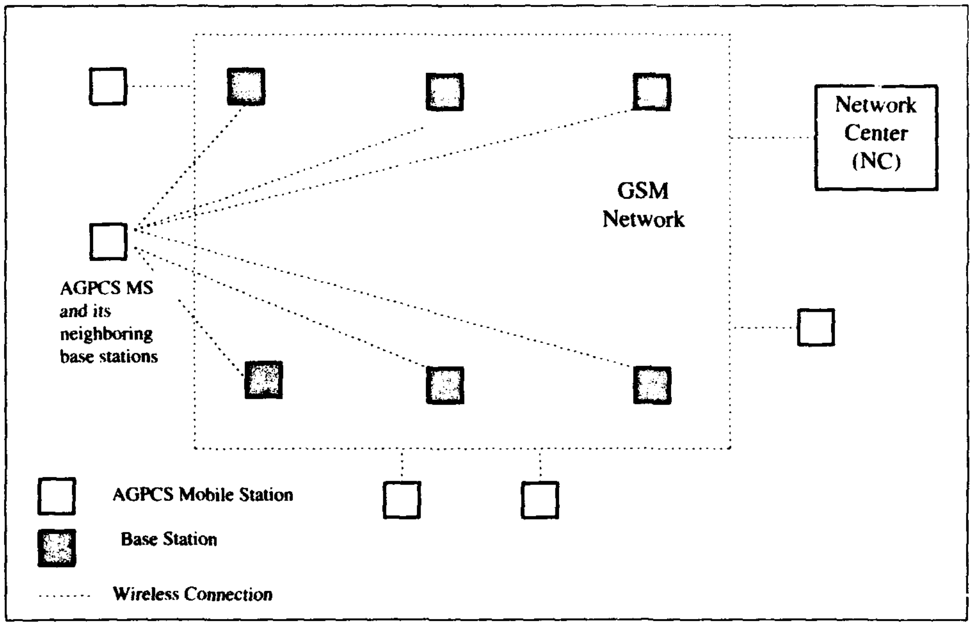


Figure 2. The AGPCS system illustration

to it. Depending on the power of the mobile computer, various degrees of intelligence and application complexity can be achieved within the AGPCS mobile station. The AGPCS mobile station performs continuous radio signal strength measurement and acquisition of measurements to estimate its position. The position is estimated by applying a combination of mathematical and statistical modeling, augmented with the use of artificial neural networks to determine position or area in which is the AGPCS mobile station. The model is based on current signal strength measurements, history of signal strength measurements, as well as some a priori knowledge of the environment. The mobile computer collects signal strength measurements from serving and up to six neighboring base stations, together with time stamps, and evaluates the distance of the mobile station from the neighbouring mobile stations. This operation is performed in real-time. Radio signal strength measurements are performed with the sampling interval between 0.5s and 1.0s. All calculated distances are used to determine an area (if there are just three distances it is a point) in which the mobile station might be. The distances of the AGPCS station from the base stations are calculated from the radio signal propagation model with parameters which are determined and subsequently changed by a training process using artificial neural networks.

Two scenarios are used from this point on. First, if the computational power of the mobile computer is sufficient, it performs further calculations, determines the estimated position, and displays it on the geographic map. It is also able to transmit its estimated position, as well as signal strength measurements, if needed, to any party in the AGPCS system, including supervisory center. This scenario leads to self-positioning and communication system, or SPCS. The SPCS is useful in applications in which the AGPCS mobile station and its user want to know current position.

In the second scenario, the mobile computer has a minimum of intelligence and input/output devices. It is used just to collect signal strength measurements, preprocess them and transmits to a network center (NC), where they are used t estimate position. A simplified version of positioning model can run on the mobile computer, and estimate the distances to the base stations, or position, which are then sent to the NC. The NC plays supervisory role in the AGPCS system. It is connected by wireless connection to the GSM. Further refinement of position can be done and the corresponding database is updated. The NC maintains data on positions of a number of mobile stations, and provides the means for presenting positions on geographic map display, but can be used for various other purposes. The NC and a number of the AGPCS mobile station make an AGPCS system. Obviously, the number of independent AGPCS systems or their architecture is not limited, because it depends only on the application requirements. Both scenarios involve transfer of messages between the AGPCS stations or between stations and the network center. This communication is performed without employing GSM voice channels. It is based on short message service (SMS) that provides exchange of short messages without using any additional interface equipment.

3. Position Estimation

Mobile station positioning is carried out by a complex combination of three types of models:

1. Geometric model based on trilateration, which gives accurate position given the distances of mobile station from the base stations.
2. Radio signal propagation model, which is of empirical character and includes many uncertain or unknown elements changing in time randomly.
3. Artificial Neural Network (ANN) model, which is used to reduce uncertainties by learning from the previous experience gained at the training station or mobile station itself.

Although the combination of these models carries certain level of redundancy, this is desirable in order to reduce the influence of the randomness, which is very high in the area of radio signal propagation and makes its modeling a very difficult task.

The main problem affecting the estimated position accu-

racy is the high uncertainty and randomness of radio signal propagation process. Using ANN modeling (Braspenning 1995, Swingler 1996) proved to have some advantages such as not requiring a priori knowledge of the relationship between dependent and independent variables. First, the model is run to learn using past and current positioning data including signal strengths and actual positions determined by more accurate GPS technique, or computer-supported and generated maps. Then, the model is validated using another set of measurements. Finally, the model is used to estimate positions. At this stage, the back propagation ANN model is used in estimating AGPCS station position. It is an adaptive multilayer feedforward network modeling technique, which is often used in non-linear system modeling and time-series prediction. Two approaches have been used in estimation of AGPCS station position:

1. Signal strengths and desired AGPCS position input-output pair. In this case, ANN learns to predict the position from relationship of signal strengths and actual position determined by GPS or obtained with high accuracy from the computer generated map.
2. Trilateration. In this approach, ANN is used to model the distance between the AGPCS station and neighboring base stations from actual measurements/positions process.

Results obtained using two described approaches are better than some other reported in literature (Song 1994). The accuracy of estimated position is better than 270m. This is still worse than those reported in (Hellebrandt 1997). However, results from this reference must be taken cautiously, because they are obtained in a fully simulated environment. All our results refer to the real system experiments with real-time estimation of position. These results have been tested on limited area of about 3 x 3 km.

The implemented model is obtained by training process on the AGPCS station itself. The AGPCS station moves in a specific area and signal strength measurements together with actual positions are recorded automatically. Then, the training process starts. It first includes analysis of various types of BP ANNs using genetic algorithm (Goldberg 1989). Genetic algorithm ranks the BP ANNs the according to values of the fitness function. Finally, the best of selected ANNs are used to implement the positioning model. In the current version of the AGPCS, the selected ANNs are implemented in software. The software implemented ANN is capable of estimating position in real-time. The whole process of signal strength measurements and measurement data preprocessing, and application of ANN's model is performed in real-time using a standard PC-compatible notebook as the mobile computer. Further details of our model will not be discussed in this paper and will be reported elsewhere.

4. Communication

The short message service (SMS) is a unique service provided in GSM that allows users to send and receive point-to-point alphanumeric messages of the length of up to 160 characters. It allows two-way messaging, store-and-forward delivery, and acknowledgment of successful delivery. This service is performed within GSM control channels and does not require the use of the voice channels. This further means that no special equipment for data service, like modems or special data cards, is needed. The SMS service operates by sending a message to the service provider message center, and it is forwarded to the destination using service provider network. A problem that can arise is the delay in delivery of the message to the destination. Although delays of this type occur occasionally, most deliveries are performed within actual real-time constraints. However, delivery cannot be guaranteed within very strict time constraints. The SMS service is used to sending position information, signal strength measurements, or to exchanging other information between the AGPCS mobile station such as telemetry measurements, or control information. An application layer of the communication protocol has been developed that provides transfers of two types of messages:

1. Short messages of the length of up to 160 characters representing a computer supported version of GSM-provided SMS service. Software in the form of a dy-

namic link library supports sending and receiving of short messages, management of message buffers and GSM mobile station local memories. This software provides a number of functions that can be used from high-level programming languages.

2. Long messages of almost arbitrary length. Because of the limited data transfer speed in GSM network, the length of the long message is practically limited to the values that depend on specific application. The software layer for long message transfers provides fragmentation and defragmentation of messages, ordering, and checks for correctness of transfer, if required.

Another important feature of the AGPCS communication is ability to create own AGPCS systems (closed networks). Once the NC is created it knows which AGPCS stations are authorised to take part in the system. The AGPCS station must first register to the NC, and report to it at the agreed time intervals or at the NC request. Otherwise, the NC may unregister it from the system. This feature is important to reduce the frequency of traffic to a minimum, because sending messages is associated with the cost. Each AGPCS station can be brought in a sort of dormant state, and awakened by the NC when needed. Also, through its own intelligence it can demand communication with the NC when predefined changes in the positions or other monitored variables occur.

The AGPCS allows yet another type of message exchange. This alternative way is using point-to-point transfers that are performed after establishing connection by dialing. The transfers are done using voice channels and actual air-time. However, this option requires additional data card plugged into the notebook computer. In the current implementation of the AGPCS system, this software relies on the complex Microsoft Telephony Application Programming Interface (TAPI). Once the connection is established, it enables guaranteed data transfers between parties at the maximum data transfer speed in GSM of 9,600 baud.

5. Implementation and Application Aspects

The first implementation of the AGPCS system uses the AGPCS stations consisting of the GSM mobile station (handset) connected to the PC-compatible notebook computer. The whole software is developed to run in the MS Windows operating environment. A small real-time kernel-like application collects signal strength measurements, and prepares them to be either sent to the network supervisory center, or used to estimate position locally. The application can decide what to transfer to the NC depending on the criteria set up by the application. The other application software is used to display the current position on the geographic map. This operation is not time critical, and software reads the current position from the file that is optionally used to log (keep track of) all positions.

Maintaining the database of the AGPCS stations including position and other information can be done in the NC. In this case the AGPCS station can be reduced to the GSM mobile station and an embedded system used just to collect measurements, perform estimate of position, and send that information to the NC. The tasks performed at the level of network center become more computationally and time consuming. The communication aspect becomes very important. It is the responsibility of application developer to keep number and frequency of transferred messages low enough to avoid bottleneck at the point of connection of the NC to the GSM.

Based on GSM, the AGPCS provides and guarantees the highest level of security through the application of encryption algorithms and frequency hopping which are fully transparent to application developers. Also, the AGPCS is internationally applicable, enabling completely new global applications without using specialised equipment or investing in expensive infrastructure.

6. Conclusions

The AGPCS technology for positioning, communication and control of mobile objects is described in this paper. The AGPCS uses standard GSM to perform all functions, making simpler combination of those tasks than in other contemporary systems. New positioning method allows to position mobile stations within 270m accuracy making technology applicable for many existing needs. Current implementations of the AGPCS mobile station and network center use PC-compatible/MS Windows hardware/software platform. They have been chosen as suitable for compatibility with many other development tools and applications. However, the AGPCS mobile station can be redesigned to the form of embedded solution. The main future research directions are further improvement of positioning model accuracy, implementation of mobile station as embedded solution, and new applications of the AGPCS technology.

References

Braspenning (1995) Braspenning P.J., Thujisman F., Weijters A.J.M.M., *Artificial Neuron Networks: An Introduction to ANN Theory and Practice*, Springer, 1995

Feher (1995) Feher K. Wireless *Digital Communications*, Prentice-Hall, 1995

Figel (1969) Figel et. al. "Vehicle location by a signal attenuation method", *IEEE Transactions on Vehicular Technology*, vol.VT-18, Nov. 1969, pp.105-109

Goldberg (1989) Goldberg, D.E.: *Genetic Algorithms: in Search, Optimization and Machine Learning*. Addison-Wesley, 1989.

Hatta (1980) Hatta M., Nagatsu T., "Mobile Location Using Signal Strength Measurements in a Cellular System", *IEEE Transactions on Vehicular Technology*, vol.VT29, pp245-252, May 1980, pp. 245-252

Kapplan (1996) Kappian E., *Understanding GPS: Principles and Applications*, Artech House, 1996

Kennemann (1994-1) Kennemann O., "Pattern Recognition by hidden Markov models for supporting handover decisions in the GSM system", *in Proc. 6th Nordic Seminar Dig. Mobile Radio Comm.*, Stockholm, Sweden, 1994, pp.195-202

Kennemann (1994-2) Kennemann O.,"Continuous location of moving GSM mobile stations by pattern recognition techniques", *in Proc. 5th Int. Symp. Personal, Indoor, Mobile, Radio Comm.*, denHaag, Holland, 1994, pp. 630-634

Ott (1977) Ott G.D., "Vehicle location in cellular mobile radio systems", *IEEE Transactions on Vehicular Technology*, vol.VT-26, Feb. 1977, pp. 43-46

Redl (1995) Redl S.H., Weber M.K., *Oliphant M.W., An Introduction to GSM*, Artech House, 1995

Scourias (1995) Scourias J., "Overview of the Global System for Mobile Communications", University of Waterloo, Canada, Technical Report, May, 1995, pp. 25

Song (1994) Song H.L., "Automatic Vehicle Location in Cellular Communication Systems", *IEEE Transactions on Vehicular Technology*, vol.43, pp902-908, Nov. 1994, pp. 902-908

Swingler (1996) Swingler K., *Applying Neural Networks*, Academic Press, 1996

Image Registration Issues for Change Detection Studies

Steven A. Israel
University of Otago
Department of Surveying
PO Box 56
Dunedin
New Zealand
israel@spheroid.otago.ac.nz

Michael R. Helfert
Southeast US Regional Climate Center
South Carolina Department of Natural Resources
1201 Main Street Suite 1100
Columbia,
South Carolina 29201
United States

Roger A. Carman
University of Otago
Department of Surveying
PO Box 56
Dunedin
New Zealand

Michael J. Duggin
State University of New York -
College of Environmental Science and Forestry
Faculty of Forest Engineering
Syracuse, New York 13210
United States

Presented at the second annual conference of GeoComputation '97 & SIRC '97, University of Otago, New Zealand, 26-29 August 1997

Abstract

Change detection studies require that all spatial information be registered to a common coordinate frame. A previous image-to-map rectification study was performed by registering pixel locations to map positions in a local coordinate frame for all images in the time series. However, the precision of this study was unable to be quantified due to the uncertainty of the map generalisation (Israel et al. 1996). A better technique is to register a single image to the coordinate frame either by using conventional survey techniques, such as GPS, or by having known camera position and orientation parameters (internal and external control). The geocoded image becomes the base map. The other images are then registered to the image base map. In this case study, we have used the North Basin of the Dead Sea as our study area. We compared our results to those found by multiple image-to-map registrations.

Introduction

Monitoring large-area temporal changes, whether human induced or naturally occurring, requires a sufficient amount of archived imagery to note the changes. Ground reference information must be available to determine the local datum for quantifying the changes that are observed. Large area monitoring is neither cheap nor easy but is required for planning and management of natural resources (Estes and Mooneyhan 1994).

Israel is exploiting the mineral resources available within the Dead Sea. To do this they are effectively draining the North Basin in evaporation ponds to the south. Israel et al. (1996) attempted to assess the changes in the sea level using manned space photography registered to a 1:250,000 scale map (REF _Ref385752423 * MERGEFORMAT Figure 1). The precision of this analysis was unable to be determined due to the uncertainty of the map generalisation. This analysis repeats the process using a geometrically corrected and georeferenced Landsat Thematic Mapper (TM) image as the registration map and to quantify mapping precision.

This project demonstrates a low cost computer processing methodology to monitor large area changes. The manned space photography is publicly available at low cost. The image area has a similar ground footprint to a SPOT scene for high spatial resolution photographs (Israel 1992). However in New Zealand, an unregistered SPOT image costs approximately three hundred times that of u s-tered manned space photography. We will show that using image-to-image registration of imagery is not only less expensive but faster and more accurate than image-to-map registration for change detection issues.

Procedures

Manned space photographs of the Dead Sea have been

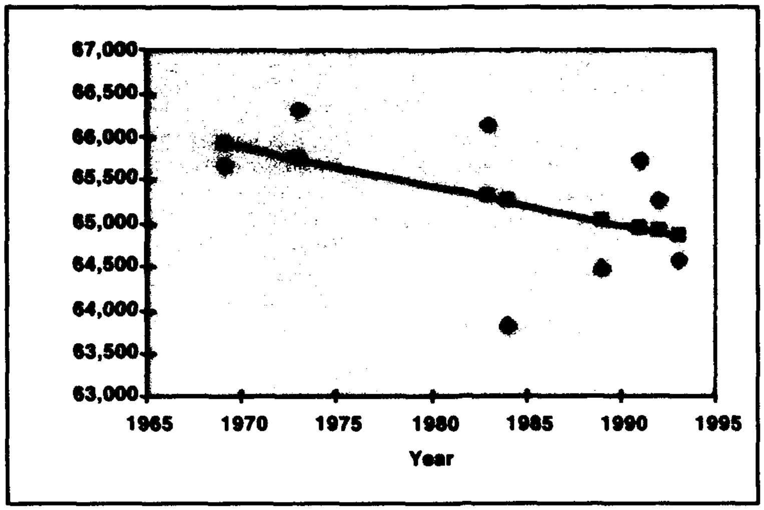


Figure 1 Decline of Dead Sea Surface Area by Year - taken from (Israel et al. 1996)
Note: ◆ indicates raw data values; and ¶ indicates linear regression line.

analysed from Apollo 9 in 1969 through to the Space Shuttle mission STS-47 September 1992. Publicly available 35 mm slides were taken from the original 70 mm format slides and tested for their suitability for analysis. Criteria for suitability were a small zenith angle of photography, a high target-to-background contrast, and a complete photographic coverage of the site and surrounding area to perform image registration (Duggin, 1990 #10).

The slides were all scanned at 600 dots per inch (dpi) and transferred to ERDAS/Imagine image analysis software for processing. 600 dpi is the highest resolution of the scanner. If the image data needed to be stored for long periods of time, then scanning resolution would have been optimised. The scanned image data was then visually inspected for usability based upon the above criteria.

Image Registration

The registration process was performed using a 1984 TM image with the standard geometric corrections (Lillesand and Kiefer 1994), as provided by United States Geological Survey (USGS). The red band of the 1984 TM image was used as it contained significant contrast between the Dead Sea and the surrounding coast and readily observable land marks for registration. The corresponding features on each manned space photographic image were registered to the TM image. Only manned space images that registered with a root mean square (RMS) error of less than 1 pixel were accepted for analysis. As each digitised manned space photograph is of different scale, the area contained by one pixel will also vary. This ground resolved cell (GRC) is a function of the acquisition parameters, film format and orbital position and orientation relative to the target area. Although the GRCs of each image pixel will vary due to the acquisition geometry, after the rectification process all image pixels contained the same linear cross section of ground projection. All rectified manned space photography images were overlaid on the TM image to visually inspect the precision of the rectification. It was found in some cases that even though the RMS error was below 1 pixel there were still obvious flaws in the rectified image. These flaws were corrected by increasing the number of registration points, especially in areas where the difference between the TM image and the manned space imagery was obvious. The image transformation was performed using the standard nearest neighbour algorithm for rectification (ERDAS 1994).

Image Analysis

Each rectified image was then analysed to establish the area of the Dead Seas North Basin in pixels. The area of each pixel in metres is known and hence the area of the North Basin on each manned space image can be calculated. Thus, the relative volume of water lost from the North Basin may be inferred.

Establishing Area-of-Interest

A single pixel was identified within the North Basin. Then, a worming function was performed to compare the digital value of the target pixel with its neighbours. The comparison is the spectral Euclidean distance. The worming function produces a vector area of interest (AOI) containing all adjacent pixels. As this is an accumulative function, each new pixel has the same function applied to its neighbouring pixels for the same range. The process continues until all pixels that are within the range, and are in contact with each other are identified. The AOI is then visually compared to the area of the North Basin. The process is repeated with different spectral distances to ensure the entire North Basin and only the North Basin is identified as one AOI. In some cases, it was not possible to identify the entire North Basin using a single AOI. In these cases, multiple AOI were identified with varying spectral Euclidean distances. These individual sub-AOIs were then merged.

Pixel Counting

The final AOI was then assessed by counting the total number of pixels and hence the total area of the North Basin. The counting procedure was repeated for an AOI with higher and lower spectral Euclidean distances.

Error Assessment of Area

The major components of error are identified. The maximum possible error due to registration is the RMS error multiplied by the total length of the major axis. In this case, the major (North-South) axis of the North Basin is multiplied by the RMS rounded to the equivalent of I pixelis GRC.

To determine the accuracy of the AOI identification some of the images were reassessed at slightly higher and lower spectral distances. This enabled us to calculate the percentage difference in total area caused by slight variations in the spectral distance. The appropriate selection of the distance defining the AOI is subjective. Recreating the AOIs

Year	Month	Image	GRC metres	Total Area pixels	Total Area hectares
1969	March	AST9 562	778	1099	66521
1982	November	STS 57-75	405	4119	67562
1983	November	S09 50 1362	343	5603	65919
1984	October	41G 120 056	687	1345	63480
1985	October	S51J 50 084	217	13371	62963
1989	October	S34 84 067	368	4800	65004
1991	April	S37 151 124	348	5401	65408
1991	June	S40 612 245	384	4393	64777
1991	June	S40 606 015	466	2835	61564
1992	March	S45 95 88	595	1766	62521
1992	September	S47 82 60	249	10474	64940

Figure 2 Results of Analysis
The results show an irregular decline in the size of the North Basin. The decline is not a linear function due to varying seasonal conditions, increases in water use, and errors in acquisition and processing. There are two areas of the analysis which can be affected by errors. The rectification processis susceptibility to errors has been minimised through using rectified images with an RMS error of less than one pixel.

Year	Month	Image	GRC	Total Area	Total Area	Max. Error Registration	Percentage Difference
			metres	pixels	hectares	hectares	
1969	March	AST9_562	778	1099	66521	4046	6
1982	November	STS 057_75	405	4119	67562	2106	3
1983	November	S09_50_1362	343	5603	65919	1784	3
1984	October	41G_120_056	687	1345	63480	3572	6
1985	October	S51J_50_084	217	13371	62963	1128	2
1989	October	S34_84_067	368	4800	65004	1914	3
1991	April	S37_151_124	348	5401	65408	1810	3
1991	June	S40_612_245	384	4393	64777	1997	3
1991	June	S40_606_015	466	2835	61564	2423	4
1992	March	S45_95_88	595	1766	62521	3094	5
1992	September	S47_82_60	249	10474	64940	1295	2

Figure 3 Registration Error Assessment

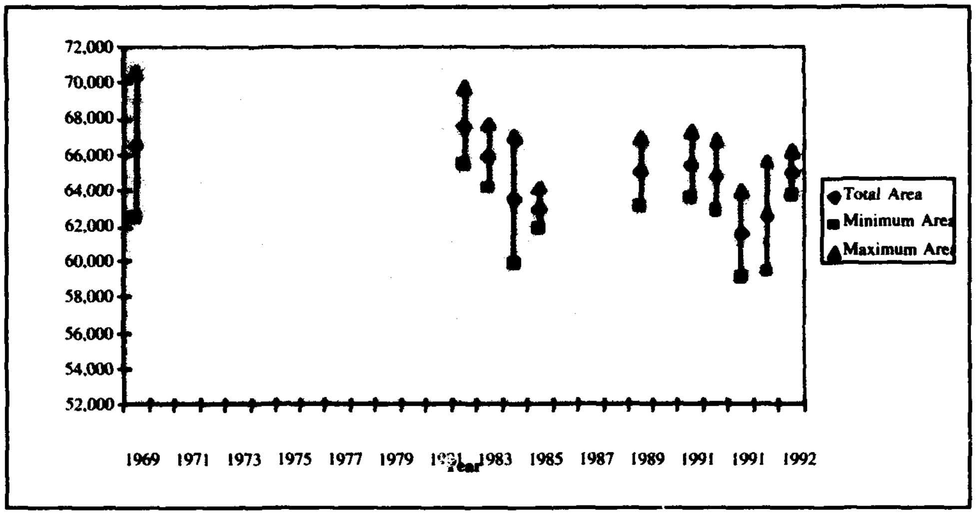


Figure 4 Rectification Error Assessment.

with higher and lower spectral differences gave us an indication of relative error due to operator subjectivity.

Results

A total of twelve images (including the TM image) were analysed. The image acquisition dates range from March 1969 through to September 1992. The results of analysis are shown in REF _Ref385749349 * MERGEFORMAT Figure 2. Relating these results to those found in REF _Ref385752423 * MERGEFORMAT Figure 1 shows little difference in the change in the area over time.

Registration Error

This means that the maximum possible error due to registration is the area of one pixel multiplied by the length (as this is larger than the width) of the North Basin (REF _Ref385751427 * MERGEFORMAT Figure 3). Given that the area we have identified as that of the North Basin (REF _Ref385749349 * MERGEFORMAT Figure 2), is correct, then the variation due to rectification error is simply that area, plus or minus the area of one pixel multiplied by the length of the North Basin (REF _Ref385749364 * MERGEFORMAT Figure 4).

Year	Month	Image	Spectral Distance	Rows	Columns	Pixel Size	Total Area	Total Area	Difference area
			digital number	pixels	pixels	metres	pixels	hectares	percentage
[illegible]	**March**	**AST9_562**	**34**	**900**	**601**	**7[illegible]8**	**[illegible]**	**66521**	
1969	March	AST9_562	29	900	601	[illegible]6	1082	65492	2
1969	March	AST9_562	39	900	601	7[illegible]8	1119	67731	2
1984	**5th August**	**TM of North Basin**	**20**	**3094**	**2045**	**28.[illegible]**	**[illegible]**	**67336**	
1984	5th August	TM of North Basin	15	3094	2045	28.5	824092	66937	1
1984	5th August	TM of North Basin	25	3094	2045	28.5	839204	68164	1
1984	**October**	**41G_120_056**	**20**	**286**	**297**	**687**	**1345**	**63480**	
1984	October	41G_120_056	15	286	297	687	1345	63480	0
1984	October	41G_120_056	25	286	297	687	1345	63480	0
1985	**October**	**S51J_50_084**	**50**	**536**	**347**	**217**	**13371**	**62963**	
1985	October	S51J_50_084	45	536	347	217	13156	61950	2
1985	October	S51J_50_084	55	536	347	217	13526	63693	1
1992	**March**	**S45_95_88**	**31**	**307**	**312**	**595**	**1766**	**62521**	
1992	March	S45_95_88	26	307	312	595	1727	61140	2
1992	March	S45_95_88	36	307	312	595	1814	64220	3
1992	**September**	**S47_82_60**	**Composite**	**554**	**453**	**249**	**10474**	**64940**	
1992	September	S47_82_60	15	554	453	249	8364	51858	20
1992	September	S47_82_60	20	554	453	249	11806	73198	13

Figure 5 AOI Error Assessment

Area-of- Interest Selection Error

As discussed earlier, the process of identifying the appropriate AOI is subjective. Once the appropriate AOI was selected the spectral distance was noted. The analysis was then repeated using spectral distances five greater and less than the original value which corresponded to + 15%. Five images were resampled to illustrate the relative errors. The results of this resampling are shown in REF _Ref385749380 * MERGEFORMAT Figure 5. The images with merged AOI are subject to the possibility of larger errors. AOI error assessment shows the percentage variation in area for each image as it is resampled with different spectral distances.

Discussion

This research quantifies the error sources associated with multidate image merging. Because the control of the registration procedure was much better than the previous attempt by Israel et al. (1996) the possibility of large errors in the image-to-map registration process was minimised (REF _Ref385749364 * MERGEFORMAT Figure 4) and consequently the error analysis was focused on the actual image analysis procedure. We also found a difficulty in pixel counting for our AOI in ERDAS/Imagine due to its approximation of pixels in an area. Consequently, we found it necessary to develop our own pixel counting software.

Our confidence in the accuracy of the data can be seen in the percentage error estimates for the samples of the data. The images with merged AOI show obvious areas of large error. This error has been somewhat exaggerated due to the error assessment being done with regards to one AOI inside the obvious boarders of the North Basin and one which is minimally outside the boarders. It was expected that images with larger GRCs would consequently show greater variability in the accuracy of total area analysis. This was not the case. It appears that the main cause of error in images is the lack of image contrast in some images between land areas and the water of the North Basin.

Conclusion

The procedures developed here may be applied to a wide range of change detection problems. Manned space photography is a low cost alternative to environmental satellite image data, and the database spans over 30 years. However, additional costs include increased registration

and computer processing time. We have shown that the cost and the processing time for these analyses can be minimised.

References

ERDAS 1994. ERDAS Field Guide, ERDAS, Atlanta, Georgia, 628 pages.

Estes, J.E. and Mooneyhan, D.W. 1994. Of Maps and Myths, Photogrammetric Engineering and Remote Sensing, 60(5): 517-524.

Israel, S.A. 1992. Manned Observations Technology Development - FY 92 Report, NASA/Johnson Space Center, Houston, Texas, JSC-26032: 34 pages.

Israel, S.A., Helfert, M.R. and Cook, A.J. 1996. Changes in the Dead Sea Over the Past 25 Years as Documented from Manned Space Photography, Environmental Geosciences, 3(1): 35-39.

Lillesand, T.M. and Kiefer, R.W. 1994. Remote Sensing and Image Interpretation, 3rd edition, John Wiley & Sons, New York, 750 pages.

Voronoi diagrams to model forest dynamics in French Guiana

F. Mercier
Laboratoire de Biométrie,
Génétique et Biologie des Populations
UMR CNRS 5558 - Université C. Bernard LYON 1
43, bd du 11 Novembre 1918,
F-69622 VILLEURBANNE Cedex

O. Baujard
Unité d'Imagerie Numérique -
Division Informatique Médicale
Hôpital Cantonal Universitaire de Genève
87, bd de la Cluse, CH-1211 GENÈVE 14

Presented at the second annual conference of GeoComputation '97 & SIRC '97, University of Otago, New Zealand, 26-29 August 1997

1. Introduction

Voronoi tessellation is an exhaustive partitioning of space in a finite set of non-overlapping continuous regions called Voronoi polygons. Such construction is defined from a finite set of distinct points. To each point (also called generator point) is associated the region of the plane which is nearer to this particular generator point than to any other (Okabe et al. 1992). In plant ecology, plant coordinates are used for generator points. Plants, being sedentary, experience the environment only in their immediate neighbourhood. Brown (1965) and Mead (1966) were the first ecologists to represent the space that closely surrounds a plant by Voronoi polygons. Brown uses Voronoi polygons as "area potentially available" to a plant, i.e. the available area for a plant to satisfy its needs in water, nutrients and light. A generator point can also be compared to a tree trunk and the associated polygon to the 'crown projection area' of the corresponding tree, used by foresters (Bouchon 1979). Voronoi tessellation (Fig. 1) gives a detailed description of the position, size and shape of individual plants in relation to the number and proximity of their contiguous neighbours. Hence, polygon features reflect local variations in plant density.

Several authors, working under carefully controlled conditions with monospecific even-aged populations, have used polygon area as a descriptive tool of spatial plant arrangement and/or as a predictive tool of plant performance (Mithen et al. 1984, Matlack & Harper 1986, Aguilera & Lauenroth 1993). Other studies, on seedling survival (Watkinson et al. 1983, Owens & Norton 1989), show that mortality is greater for plants of smallest polygon area.

In this article, we propose another use of Voronoi diagram in constructing a spatiotemporal model to study plant population dynamics. We analyse in particular the influence of various recruitment processes on the spatial patterns of a Guianan forest stand. First, an initial model containing a random recruitment process provides results on age, population size and spatial pattern dynamics. Then, we conceive a second Voronoi model including canopy openings and recruitment processes in gaps. With this second model, we focuse our

attention on the changes in spatial patterns through time according to the opening rate and gap area distribution.

2. Voronoi models

2.1. Concepts and implementation

A Voronoi tessellation can be defined as follows : let $P_1, P_2, ..., P_N$ be a finite number of distinct points in the plane, the region associated to P_N is the set T_N defined by:

$T_N=\{x \mid d(x,P_n)\leq d(x,P_m) \text{ for all } m\neq n\}$

where d is the euclidean distance (Okabe et al., 1992).

To construct the Voronoi tessellation, we use the algorithm proposed by Green & Sibson (1978), revised by Bowyer (1981) and Bertin (1994). It consists of an incremental method of adding generator points one at a time

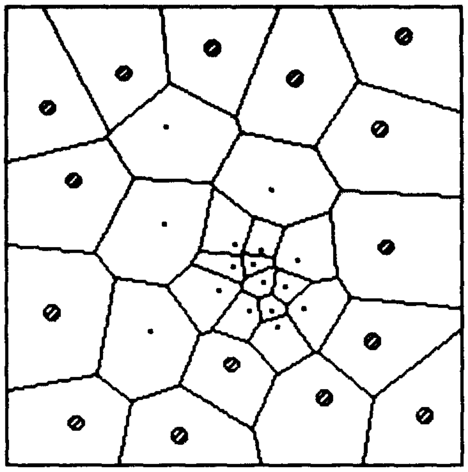


Figure 1 : Voronoi tessellation where polygons containing small hatched circles are marginals.

in the sampling window until the tessellation is complete. The injection of a new generator point modifies local contiguities. Such an algorithm uses lists of generator points and vertices, and is computationally efficient (resolution time in O(n)).

Certain polygons, called marginal, are partially determined by sampling window boundaries. Such marginal polygons are not representative of the population and should be excluded from any analysis. To select marginal polygons, we use the algorithm proposed by Kenkel et al. (1989a).

To use Voronoi tessellation for spatiotemporal models, some generator points are inserted and others are suppressed from the tessellation through time steps, according to rules for recruitment (arrival of saplings in the stand)

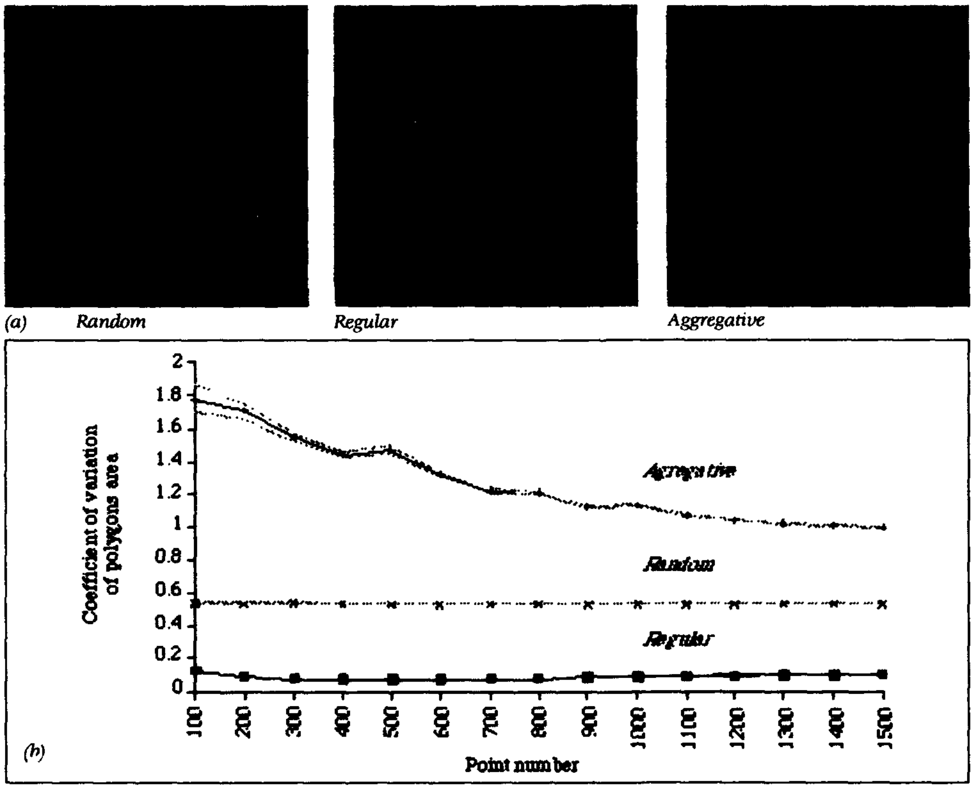


Figure 2 : (a) Voronoi tessellation constructed with 618 points determined from three different point processes. (b) Coefficient of variation of polygon area with respect to the number of points, to detect the spatial pattern. Curves are bounded by confidence intervals obtained from Monte Carlo simulations of aggregative (Neyman-Scott), random (Poisson) and regular (randomized periodic) point processes.

and mortality (removal of trees). Thus, insertion and suppression of points will induce local modifications in the simulated forest stand spatial patterns. However, recruit number and tree mortality are functions of the total population size. Thus, while the population size dynamics is managed at a global level, the changes in spatial pattern through time arise from local events. Furthermore, plants can change their internal state (such as age or diameter) according to a growth model. In this article, age is incremented at each time step but the growth process is not included.

A preliminary analysis of Voronoi polygon properties led us to prefer the coefficient of variation of polygon area (CV) as the most simple and efficient variable for describing the spatial pattern of generator points (Fig. 2) (see also Vincent et al. 1976, Upton & Fingelton 1985, Hutchings & Discombe 1986, Lorz 1990, Marcelpoil & Usson 1992).

2.2. Random recruitment hypothesis

A first model has been conceived for analysing the behaviour of Voronoi polygons used in spatiotemporal models. At the initial time, a Voronoi diagram is constructed with 618 points corresponding to the mean density of trees (with a dbh(*) >10cm) observed on 1 ha at the Paracou experimental site (Schmitt & Bariteau 1990) in French Guiana (5°15'N, 52°55'W) between 1984 and 1994 (Fig. 3). The initial points are randomly distributed following a Poisson point process, in accordance with the spatial pattern of trees observed on field data. At each step, r individuals are recruited and m trees are removed, such as :

$N(t+1) = N(t) + r - m,$

with $r = Bin(R, N(t))$ and $m = Bin(M, N(t))$

where Bin(n, p) refers to the binomial distribution with n, the number of trials and p, the success probability. The symbol R represents the recruitment rate, M, the mortality rate and N(t), the population size at time t. The coordinates of the recruits are determined from a Poisson point process and each individual has the same probability of being eliminated. As trees are recruited at dbh = 10 cm, one time step equals the necessary time to reach such a diameter i.e. approximately 10 years.

(*) dbh = diameter at breast height is the more common measure of tree size.

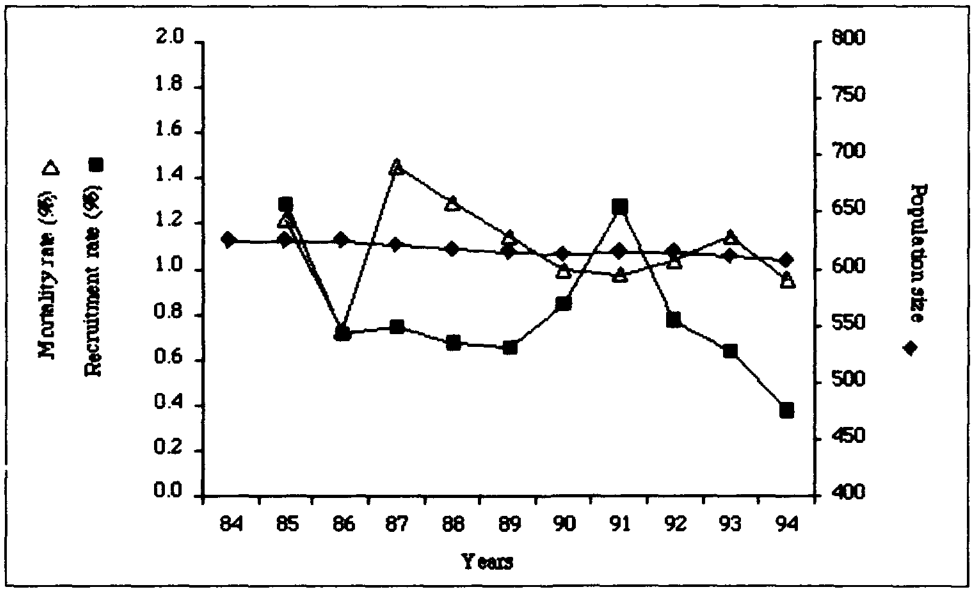


Figure 3 : General characteristics of the forest dynamics at Paracou experimental site (French Giuana) between 1984 and 1994. Recruitment and mortality rates are expressed in % of the population size.

During the simulation procedure, we test several values for R and M, including the extreme ones observed on the Paracou site. The simulations are realized for different initial spatial patterns : complete spatial randomness (Poisson point process), Neyman-Scott aggregative process (see Stoyan et al. 1995 for a review) and "muddled" periodic spatial pattern. For each set of parameters, 30 simulations are performed in order to obtain statistically valid results. The output variables i.e. population size, age and CV, are observed on 200 time steps.

2.3. Results

The system is obviously sustainable when M approaches R but a small difference between the values of R and M lead to a fast deviation from equilibrium (Fig. 4).

On Paracou station, we observed the rates R = 0.89 % N/year and M = 1.05 % N/year. When the model runs with these values, the population dynamics is unsteady and the simulated forest stand perishes after 122.3 steps, i.e. 1223 years.

However, this result is founded on the unlikely hypothesis of persistence of the values of R and M over several centuries. As the population size dynamics seems sensitive to small differences between recruitment and mortality rates, the next model will be based on the hypothesis of a steady state of the forest stand, such that M = R = Bin(p, N(t)), where p represents both recruitment and mortality rates.

The age distribution of the trees becomes stable between 25 and 50 steps depending on the values of R and M (Fig. 5).

Whether the initial point process is random, aggregative or regular, the spatial pattern becomes random after 20 time steps (Fig. 5). This phenomenon follows from the random choices of the coordinates of recruits and the identity of trees to suppress.

3. *Gap effects on forest stand spatial patterns*

3.1. Canopy gap modelling

On average, 1% of the forest canopy is annually opened by treefalls and branchfalls. In these canopy gaps (Brokaw 1982)

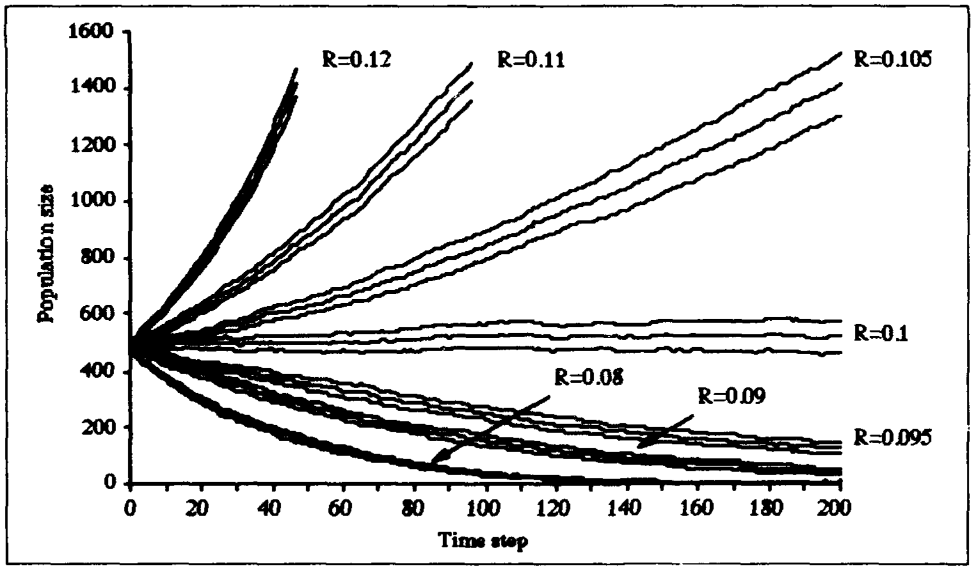


Figure 4 : Influence of the difference between recruitment and mortality rate on the population size dynamics. R represents the recruitment rate (%/step i.e. %/10 years). M, the mortality rate, equals 0.1 N/step. Curves are bounded by confidence intervals obtained with 30 simulations.

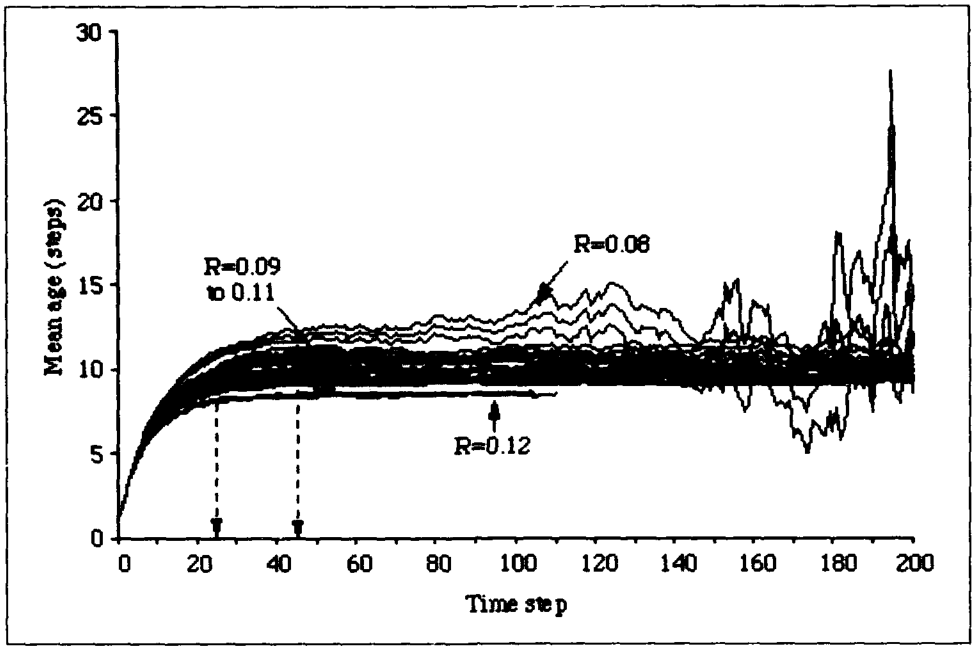


Figure 5 : Influence of the difference between R and M on the mean age of the forest stand, with M equals 0.1 N. Curves are bounded by confidence intervals obtained with 30 simulations.

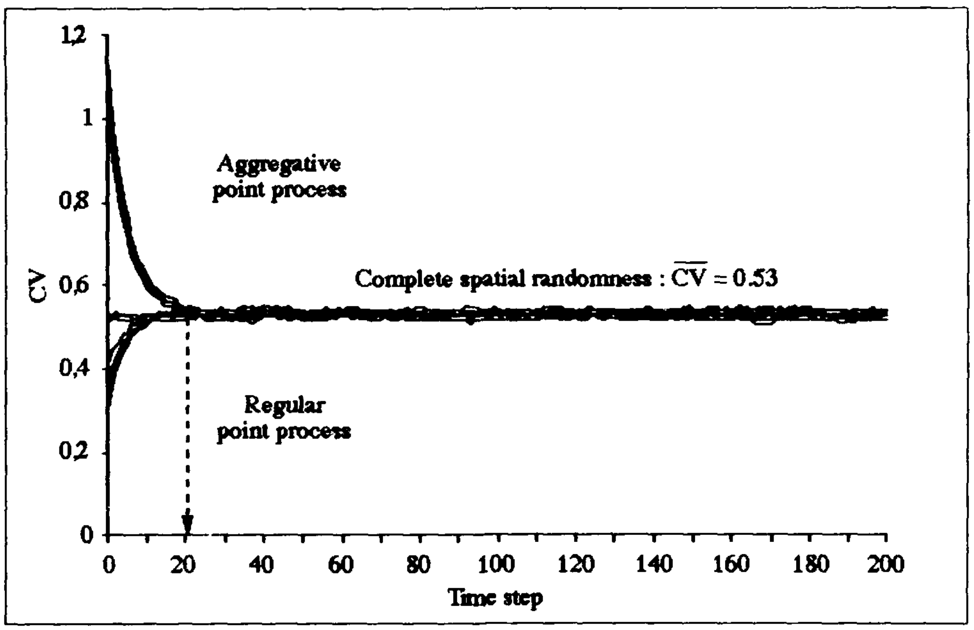


Figure 6 : Coefficient of variation of polygon areas (CV) during simulations of a forest stand dynamics with Voronoi models initialized with different spatial patterns of trees. Curves are bounded by confidence intervals obtained with 30 simulations.

new patches of vegetation start to grow which later will form the forest canopy. Gaps in the canopy increase light levels and modify other characteristics of the environment (Denslow 1987, Brown 1993) sufficiently to influence the dynamics of the tree population (Pickett & White 1985, Platt & Strong 1989, Van der Meer 1995). Numerous seedlings establish themself in these openings, inducing a clumped spatial pattern (Armesto et al. 1986).

Our aim is to estimate the aggregation intensity obtained in a simulated forest stand where canopy gaps appear. Consequences of different opening rate, gap area distribution and initial spatial stand pattern on the forest dynamics are analysed with regards to the age, population size and CV changes through time.

To include the canopy opening process in the Voronoi model, we determine at each time step, a total opened area, tg(t), such as :

tg(t) = Norm (mg, vg) x A,

where Norm refers to the normal law, mg, the mean opening rate, vg, the variance of opening rate and A, the total area of the study plot. The total opened area is spread over several gaps whose areas, sg(i), verify $\sum_i sg(i) = tg(t)$.

Values of sg(i) are samples of a Gamma law fitted on the size distribution of field observed gaps (Fig. 7).

Openings are assumed circular and their centers are randomly located in the plot. Trees located in gaps are eliminated (gap mortality process) and the openings are immediately filled by recruits. The number of recruits, r, is proportional to the gap area, as the recruitment density approaches stand density. While all recruits appear in gaps, only 52% of trees die through canopy openings; the rest (48%) are dead standing trees (Durrieu de Madron 1993). Hence, the population dynamics is expressed as :

N(t+1) = N(t) + r - (Mg + Mds),

with Mg~{sg(i)}, Mds = (0.48/0.52) Mg and $r = \sum_i (N(t)/A)\, sg(i)$,

where Mg, the number of fallen trees during opening of the canopy, depends on the size of the gap and Mds, repre-

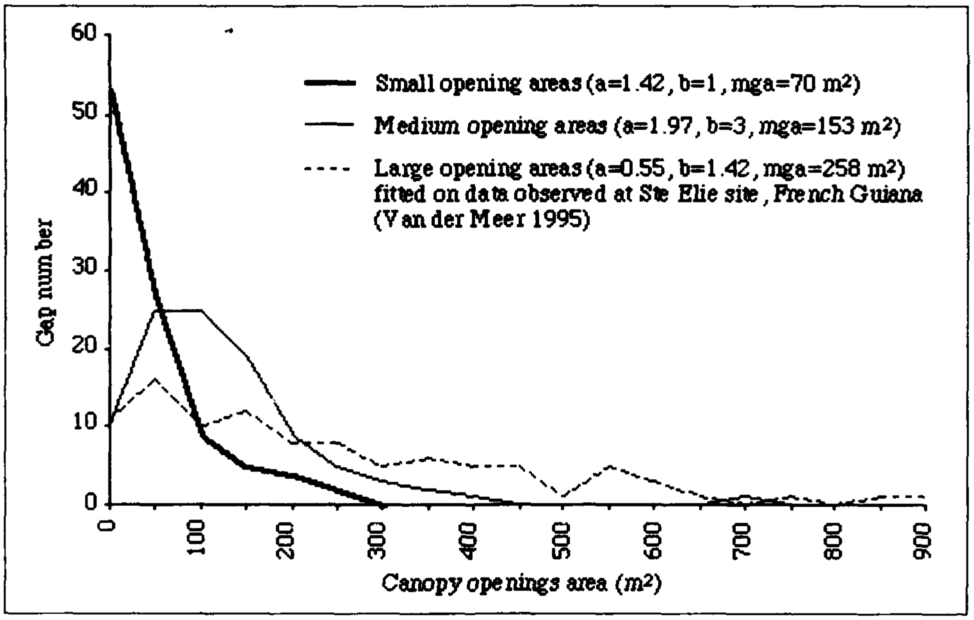


Figure 7 : Gamma laws fitted on observed gap distributions in French Guiana (Van der Meer 1995) or generated artificially. The parameters a and b represent respectively the scale and shape parameters of the Gamma law. With respect to the mean value of a Gamma law, mga is the expected mean gap area mga = (100 b/a)

sents dead standing trees.

The combination of three values of tg, three couples a, b and three initial spatial patterns allow us to simulate various disturbance modes of the neotropical rain forest (Table 1). As previously, 30 simulations are realized on 200 steps for each set of parameters. The observed output variables are population size and CV, polygon area and tree age distributions, plus some information on gap characteristics (number of gaps, mean and variance of their areas).

3.2. Results

Changes in opening rate and Gamma function parameters (a and b) imply variations in the gap numbers (Table 2). Gap number increases with tg but the mean gap area remains equal to 159.2 m^2 (SD = 6.8). Thus, turnover rate(*) increases and, consequently, mean age decreases because the total opened area of the forest stand increases with tg (Fig. 8a). The necessary time to reach a stationary mean age (the "transient regimen") decreases as tg increases (Fig. 8a).

Though the gap number decreases when the mean area of the simulated gap increases (Table 2), the age distribution of the forest stand remains unchanged because the total opened area is the same however it is split into individual gaps (Fig. 8b).

The coefficient of variation of polygon areas varies according to the opening rate but also to the gap area distribution. The box plots of CV illustrate these differences (Fig. 9). We conclude that gap dynamics plays an important role in generating a tendency to aggregation in the spatial pattern of the forest stands.

Finally, we analyse the effect of the initial spatial pattern of trees on changes in CV with time, when the opening rate equals 10 %/step and the gap area distribution corresponds to an intermediate case (a=1.97, b=3.00). The transient regimen is shorter than the observed one in the reference model, and the CV mean is higher (0.75) (Fig. 10). Thus, the introduction of the canopy opening mechanisms seems to ...he aggregativity of the forest stand.

4. Dicussion and perspectives

The study of tropical forest dynamics is based on the analysis of three closely linked elements : first, the population size influenced by recruitment and mortality mechanisms and secondly, the diameter distribution or basal area which depends on the growth processes. The last element is the spatial distribution of the forest stand. Usually, we accept the following sequence :

clumped juveniles -> random adults -> regular old adults

(*) Turnover rate = number of years it takes to cover a unit area of forest with gaps, using the average area annually affected by gaps (Van der Meer 1995).

Table 1 : Simulated gaps features

Initial spatial point pattern	tg (%/step)	a	b
Poisson point process	5	1.42	1.00
"Muddled" periodic point process	10	1.97	3.00
Neyman-Scott point process	20	0.55	1.42

Table 2 : Gap numbers obtained from different opening rates or gap area distributions

Opening rate (%/step)	Mean gap number (SD)	Gamma parameters (a, b)	Mean gap area (SD) (m^2)	Mean gap number (SD)
for mga=159.2		for tg=10		
5	3.9 (0.19)	(1.42, 1.00)	75.5 (4.3)	15.1 (0.75)
10	7.2(0.31)	(1.97, 3.00)	159.2 (6.8)	7.2 (0.31)
20	13.8 (0.69)	(0.55, 1.42)	298.0 (20.2)	4.8 (0.34)

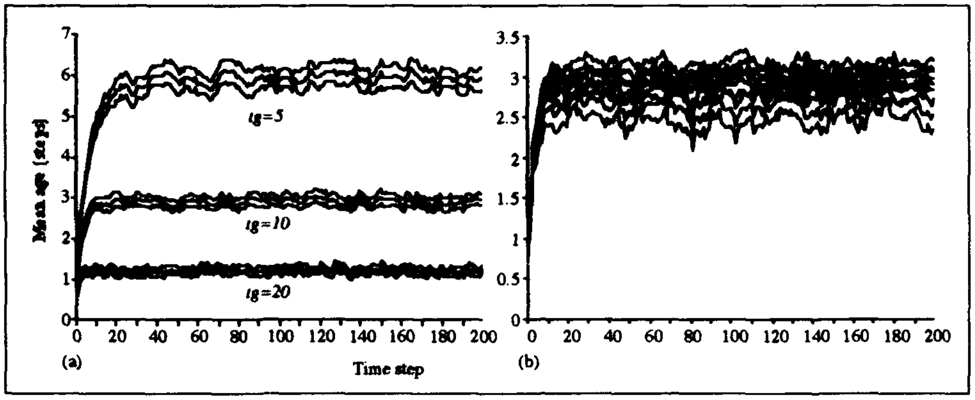


Figure 8 : Mean age of simulated forest stands. Curves are bounded by confidence intervals obtained with 30 simulations. (a) Effect of the opening rate when the mean gap area (mga) equals the typical value of 159.2 m². *(b) Effect of the gap area distribution when the opening rate (tg) equals 10%/step.*

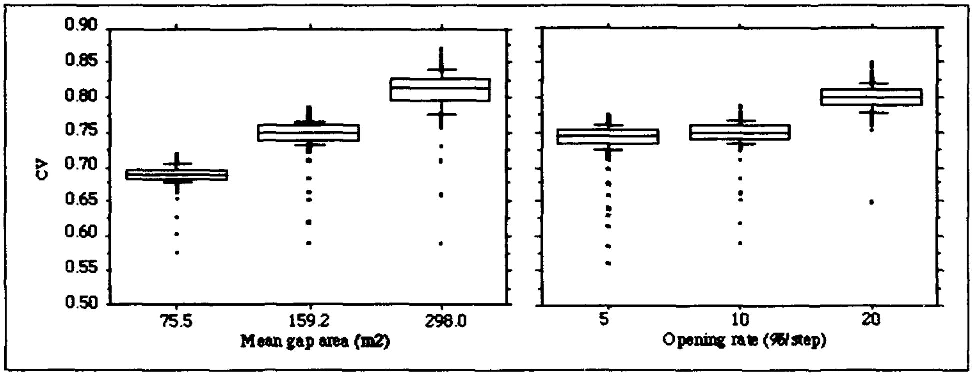


Figure 9 : The coefficient of variation of polygon areas as functions of (a) opening rate and (b) gap area distribution.

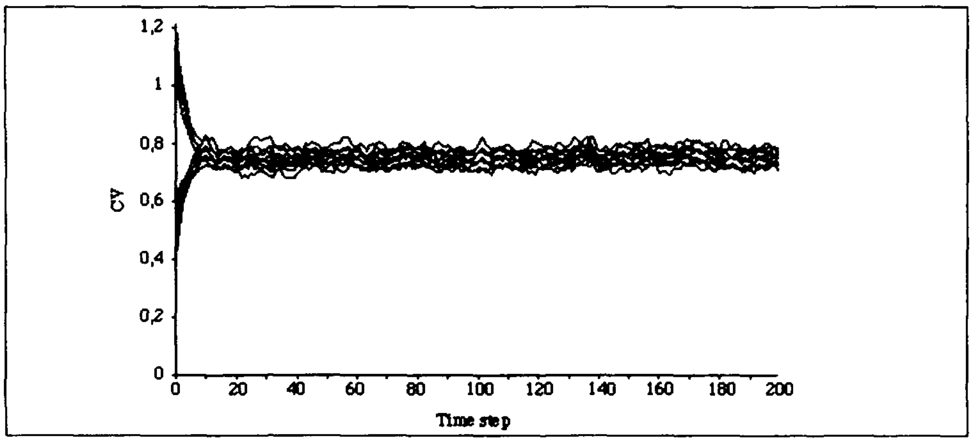


Figure 10 : CV dynamics when the opening rate equals 10%/step and the disturbance mode corresponds to a intermediate case (a = 1.97, b = 3.00). Curves are bounded by confidence intervals obtained with 30 simulations.

to describe the changes in the spatial patterns of a forest plot with time (Kenkel 1988, Gavrikov & Stoyan 1995). This sequence results from a massive recruitment of seedlings which induce the formation of aggregates in zones favourable to sapling establishment. The high density characterizing these zones then triggers a self-thinning process in the plot due to the competitive interactions for resources. So, as time advances, we observe a repulsion phenomenon between individuals, leading to a regular distribution of the forest stand.

This theoretical scenario is not always verified and the underlying mechanisms are not always known. Typically, in the above time sequence, gap influence is not considered. The studies realized on French Guiana forests suggest the intervention of at least four factors in the spatial structuration of the forest stand : competition between individual trees, seed dispersal, soil features and canopy openings.

As canopy openings are propitious zones for recruitment, we test the hypothesis of an increased aggregation intensity in disturbed forest. The results give an average CV=0.75 in disturbed simulated forests vs. 0.53 in undisturbed ones, leading us to conclude on an aggregative effect of canopy openings. In addition, the aggregation rate increases with opening rate. The same trends appears when the mean gap area increases from 75.5 m^2 (SD = 4.3) to 298.0 m^2 (SD = 20.2). Though these results were expected, the observed values of CV for the plots of primary forest in Paracou station remain near random at 0.53 while the mean opening rate equals 1% per ha and per year.

Obviously, it is unrealistic to consider canopy gaps as the unique factor managing the spatial pattern changes in neotropical forests. Consequently, inclusion of the competitive interactions will be the next step in the development of our Voronoi model of forest dynamics. Competition could provoke a self-thinning process to counterbalance the aggregative effect of gaps. The individual based and spatially explicit models applied to forest dynamics study use various expressions of competition in their growth submodel. Several authors suggest using the area of Voronoi polygons as a competition index. However, Kenkel et al. (1989) and Walden et al. (1990), exploring this form of competition index on forest stands explain less than 30% of the growth variation of studied trees. Indeed, one drawback of this approach is ignoring individual tree size, since the polygon area depends solely on the positions of neighbouring points.

To conclude, the model presented in this article introduces a new way to study the forest dynamics with spatiotemporal models (Czàran & Bartha 1992). Voronoi diagrams offer the opportunity to analyse simultaneously the spatial pattern of the forest stand and the local competition pressure occuring between trees. An example of such a Voronoi forest model led us to highlight the determinant role of gaps in generating an aggregative spatial pattern of trees. To analyse the effect of interindividual competition on spatial pattern and diameter distribution, in further work we will introduce a growth process in our model with a competition index that considers both polygon area, size of the corresponding tree and size of its neighbours.

Acknowledgements

We would like to thank A. Pavé and C. Gautier, both professors at Université Lyon I for supporting this project. We thank L. Houde, E. Bertin and C. Deleuze for their constructive remarks on Voronoi diagrams and modelling strategy. Finally L. Blanc, E. Desouhant, R. Grantham and M.A. Moravie are acknowledged for their helpful comments on the manuscript. F. Mercier received a grant from the Ministère de l'Enseignement Supérieur et de la Recherche, and complementary financial support was attributed by CIRAD-Forêt to insure working expenses in the course of his PhD thesis.

References

Aguilera M.O. & Lauenroth W.K. (1993). Neighborhood interactions in a natural population of the perennial bunchgrass Bouteloua gracilis. Oecologia, 94 : 595-602.

Armesto J.J., Mitchell J.D. & Villagran C. (1986). A comparison of spatial patterns of trees in some tropical and temperate forests. Biotropica, 18 : 1-11.

Bertin E. (1994). Diagrammes de Voronoi 2D et 3D : Applications en analyse d'images. Thèse de Doctorat. Université J. Fournier (Grenoble).

Bouchon J. (1979). Structure des peuplements forestiers. Annales des Sciences Forestières, 36 : 175-209.

Bowyer A. (1981). Computing Dirichlet tessellations. The Computer Journal, 24 : 162-166.

Brokaw N.V.L. (1982). The definition of tropical gap and its effect on measures of forest dynamics. Biotropica, 14 : 158-160.

Brown G.S. (1965). Point density in stems per acre. New Zealand Forestry Service Research Notes, 38 : 1-11.

Brown N. (1993). The implications of climate and gap microclimate for seedling growth conditions in a Bornean lowland rain forest. Journal of Tropical Ecology, 9 : 153- 168.

Cz‡ran T. & Bartha S. (1992). Spatiotemporal dynamic models of plant populations and communities. Trends in Ecology and Evolution, 7 : 38-42

Denslow J.S. (1987). Tropical rainforest gaps and tree species diversity. Annual Review of Ecology and Systematics, 18 : 431-451.

Durrieu de Madron L. (1993). Mortalité, chablis et rôle des trouées dans la sylvigenèse avant et après exploitation sur le dispositif d'étude sylvicole de Paracou - Guyane française. Thèse de Doctorat. ENGREF.

Gavrikov V.L. & Stoyan D. (1995). The use of marked point processes in ecological and environmental forest studies. Environmental and Ecological Statistics, 2 : 331-344.

Green P.J. & Sibson R. (1978). Computing Dirichlet tessellations in the plane. The Computer Journal, 21 : 168-173.

Hutchings M.J. & Discombe R.J. (1986). The detection of spatial pattern in plant populations. Journal of Biogeography, 13 : 225-236.

Kenkel N.C. (1988). Pattern of self-thinning in jack pine : testing the random mortality hypothesis. Ecology, 69 : 1017-1024.

Kenkel N.C., Hoskins J.A. & Hoskins W.D. (1989a). Edge effects in the use of area polygons to study competition. Ecology, 70 : 272-274.

Kenkel N.C., Hoskins J.A. & Hoskins W.D. (1989b). Local competition in a naturally established jack pine stand. Canadian Journal of Botany, 67 : 2630-2635.

Lorz U. (1990). Cell-area distribution of planar section of spatial Voronoi mosaics. Materials Characterization, 25 : 297-309.

Marcelpoil R. & Usson Y. (1992). Methods for study of cellular sociology : Voronoi diagrams and parametrization of the spatial relationships. Journal of Theoretical Biology, 154 : 359-369.

Matlack G.R. & Harper J.L. (1986). Spatial distribution and the performance of individual plants in a natural population of Silene dioica. Oecologia, 70 : 121-127.

Mead R. (1966). A relationship between individual plant-spacing and yield. Annals of Botany, 30 : 301-309.

Mithen R., Harper J.L. & Weiner J. (1984). Growth and mortality of individual plants as a function of available area. Oecologia, 62 : 57-60.

Okabe A., Boots B. & Sugihara K. (1992). Spatial tessellations concepts and applications of Voronoi diagrams. John Wiley & Sons.

Owens M.K. & Norton B.E. (1989). The impact of 'available area' on Artemisia tridentata seedling dynamics. Vegetatio, 82 : 155-162.

Pickett S.T.A. & White P.S. (1985). The ecology of natural disturbance and patch dynamics. Academic Press.

Platt W.J. & Strong D.R. (1989). Special feature : Gaps in forest ecology. Ecology, 70 : 535-576.

Schmitt L. & Bariteau M. (1990). Etude de la croissance et de la régénération naturelle. Dispositif de Paracou. Bois et Forêts des Tropiques, 220 : 3-23.

Stoyan D., Kendall W.S. & Mecke J. (1995). Stochastic geometry and its applications - Second Edition. John Wiley & Sons.

Upton G. & Fingleton B. (1985). Spatial data analysis by example. Volume I : Point pattern and quantitative data. John Wiley & Sons.

Van der Meer P.J. (1995). Canopy dynamics of a tropical rain forest in French Guiana. Thesis, Wageningen Agricultural University, The Netherlands.

Vincent P.J., Haworth J.M., Griffiths J.G. & Collins R. (1976). The detection of randomness in plant patterns. Journal of Biogeography, 3 : 373-380.

Welden C.W., Slauson W.L. & Ward R.T. (1990). Spatial pattern and interference in pinon-juniper woodlands of northwest Colorado. Great Basin Naturalist, 50 : 313-319.

Management of Gridded Climate Data for National Scale Integrated Assessment Models

G. C. Sims, R. A. Warrick, G. J. Kenny and W. Ye.
Centre for Environmental and Resource Studies (CEARS)
University of Waikato, PB 3105, Hamilton, New Zealand
g.sims@waikato.ac.nz

Presented at the second annual conference of GeoComputation '97 & SIRC '97, University of Otago, New Zealand, 26-29 August 1997

Abstract

Computer model analyses of climate change impacts are data intensive due to the spatial and temporal dimensions over which climate operates. Data intensity proves a major constraint in the design of such climate models. For policy oriented climate models this constraint proves critical, given the lower specification computer hardware readily available to decision makers. This paper discusses the use of spatial data orderings in combination with run-length encoding to spatially compress climate data. Experiments have been conducted which test the application of various data ordering schemes to the storage of climate data for New Zealand, Australia and Bangladesh. The results of these experiments are presented.

1. Introduction

Under the Framework Convention on Climate Change, signatory parties have an obligation to report to the Conference of the Parties regarding their vulnerability and adaptive capacity to climate change. This places reporting countries in an awkward position; policy makers are advised that the greenhouse effect is real, and probably already occurring, but they often have little quantitative information on the impacts of global warming on which to base their assessments. To make informed decisions, policy makers need tools which enable them to estimate the implications of climate change over a wide range of policy options, and which can provide a concise overview of the uncertainties surrounding global climate change (Hulme *et al.*, 1994; Dowlatabadi and Morgan, 1993). Importantly, this requires the consideration of the spatio-temporal impacts of climate variability and change.

The most efficient way of dealing with climate impacts over time and space is through the use of computer models. However, the development of computer models for climate impact assessment is fraught with difficulty. Due to the spatial and temporal nature of the analyses, such climate impact models usually process data over at least two dimensional space, and thus tend to be data intensive. Typically, information systems store and manipulate one dimensional data. Data, therefore, proves to be a bottle-neck in many climate impact models, requiring significant amounts of storage space and fast computer hardware (notably disk drives and processors). As such, research is necessary to improve the design and implementation of data structures and algorithms for the management of spatially referenced climate data. This paper examines techniques for the storage of spatial climate data. Attention is focused on the use of various data orderings in combination with run-length encoding (RLE) to reduce storage requirements.

2. Integrated Assessments Models — the context

As noted, there is an immediate need for policy decisions on how to prevent or adapt to climate change. For this, information on climate change is fundamental. However, the most scientifically advanced climate models, general circulation models (GCMs), are too computationally demanding for such purposes, generally requiring large

amounts of computer processing power and time. Essentially, the complexity of such models makes them more suited to scientific analyses, rather than for direct use in policy or impact analysis which tend to require multiple model runs. Additionally, the spatial resolution of GCMs is often too low to prove of any real benefit for national or local scale policy or impact analysis. Finally, GCMs themselves say little about biophysical and socio-economic impacts or mitigation and adapation options.

To overcome this methodological gap, a new class of *integrated assessment models* (IAM) has evolved (Weyant *et al.*, 1996). Such systems combine climate, environmental and socio-economic impact models in order to provide the flexibility to evaluate the effects of climate change and variability. Often, these systems integrate subjective expert judgement about poorly understood parts of the problem, with formal analytical treatment of the well understood parts (Dowlatabadi and Morgan, 1993). These IAMs typically attempt to capture the most salient features of more advanced climate models in a reduced-form, or as results generated off-line and used as model input data. Modularity, inherent in IAMs, ensures the software is readily updated to reflect scientific advances. The most comprehensive and complex versions of IAMs are the highly aggregated global-scale IAMs (eg. IMAGE: Alcamo *et al.*, 1994)

At a national scale, simpler integrated models are being developed for New Zealand (CLIMPACTS), Bangladesh (BD-CLIM), and Australia (OzClim) (Kenny *et al.*, 1995; Warrick *et al.*, 1996). The purpose of these models is to examine the spatial impacts and sensitivities of various sectors (in New Zealand, for example, pastoral, horticultural and arable cropping sectors are examined) to climate variability and change. The models can be viewed as a graphic user interface that provides a structured route through a collection of climate and sectoral impact models. In essence, they operate by coupling a simple global climate model (MAGICC - Model for the Assessment of Greenhouse gas Induced Climate Change, see Osborn and Wigley (1994); Wigley and Raper, (1992); Wigley and Raper, (1993); Wigley, (1993); Hulme *et al.*, (1994)) with a regional climate change model to generate scenarios (raster images of climate variables) in five year increments to the year 2100. These images are, in turn, used as input to the sectoral impact models. For a more complete description of this methodology see Kenny *et al.* (1995) and Warrick *et al.* (1996).

In designing such national-scale integrated models, it is important to consider computational efficiency. A policy-oriented tool should allow multiple experiments to be undertaken quickly, thus allowing sensitivity analysis of various model inputs and assumptions. However, many integrated models are designed to run on desktop computers. The reduced processing power, memory and secondary storage (disk space) of desk top computers is a determinant of the spatial and temporal resolutions at which the software operates, as well as of the scientific complexity of the model. Thus, there is a need to increase computational efficiency. Some techniques, researched in the context of national-scale IAM development, are discussed below.

3. Spatial data-structures

3.1 Run-Length Encoding

Many features which are mapped change gradually over space. If such a feature is mapped in a raster format, there is a probability that neighbouring cells will have the same attribute value. As such, raster maps and images generally have some degree of homogeneity (Bell *et al.*, 1988). The degree of homogeneity depends on important factors such as the spatial variability of the feature and the resolution it is being mapped at. Figure 1 illustrates a simple map, and a possible raster representation of this map. Although this is an example, it illustrates that often cells in a raster image have the same value as a neighbouring cell.

If the grid is read from left to right (row order) it is evident that there is repetition of data. Table 1 illustrates the one dimensional row order representation of the above grid within a file. As can be seen, there is considerable redundancy in the file due to repetition of values. This is common in spatial data with some degree of homogeneity. Run-length encoding (RLE) takes advantage of homogene-

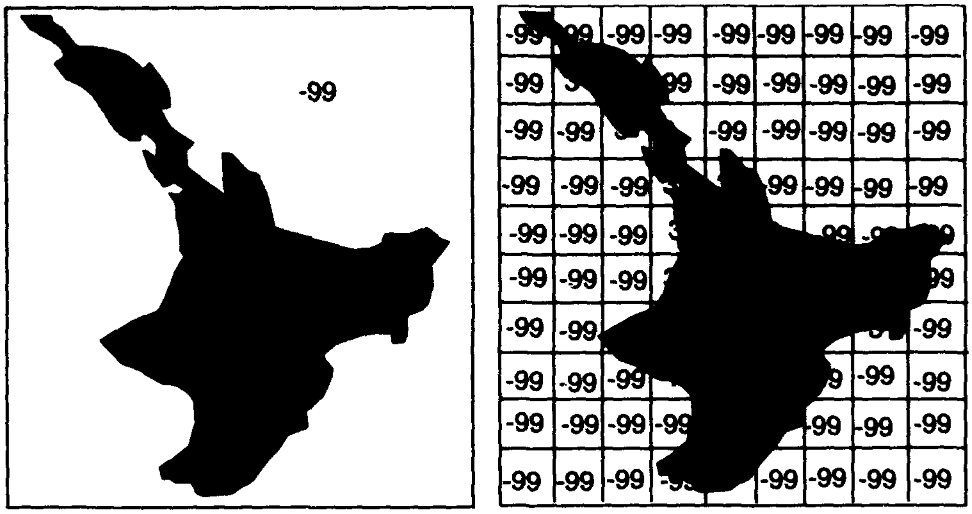


Figure 1, Raster Representation

-99 -99 -99 -99 -99 -99 -99 -99 -99 -99 3 3 -99 -99 -99 -99 -99 -99 -99 -99 3 3 -99 -99 -99.....

Table 1, Row Order File Structure

-99:10 3:2 -99:8 3:2 -99:8 3:2.....

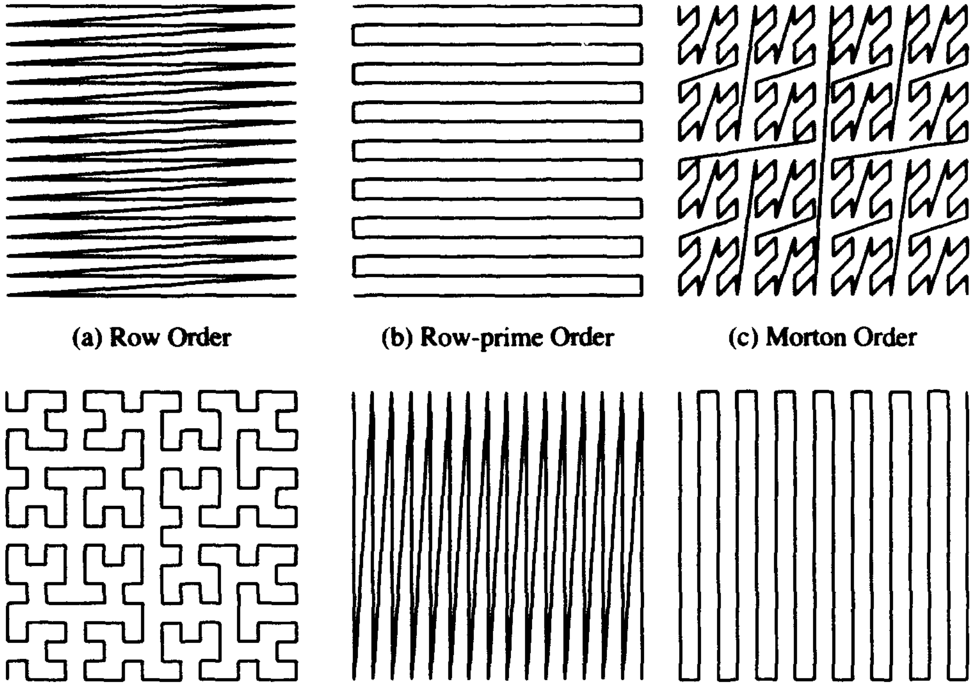
Table 2, Row Order File Structure with Run-Length Encoding

ity in data to reduce the amount of disk space necessary to store the data (Eastman, 1992a; Eastman, 1992b; Holroyd and Bell, 1992; Goodchild and Grandfield, 1983; Abel and Mark, 1990). When writing a raster to file using RLE, repetitions of values are recognised. Instead of writing each individual value to the file (as in the above example), as each repetition of values is encountered, the value is written once along with the number of repetitions of it (the run-length). A record can be eliminated from the file whenever a cell has the same value as the cell previously processed. Table 2 illustrates the row order representation of the above grid using RLE.

3.2 Data Ordering

A raster image can occupy different amounts of storage depending on how it is structured and ordered. To benefit more from RLE and reduce storage requirements, homogeneity can be increased by using different data orderings. Geographic data are essentially two (or more) dimensional, whereas computer storage and processing are essentially one dimensional (Mark, 1986). No linear (one dimensional) sequence can preserve, and therefore benefit from, all spatial properties of geographic data (Mark, 1986). Using RLE, longer run-lengths will result in less storage requirements. Intuitively then, one would expect orderings which attempt to best preserve spatial relationships to increase run-lengths and reduce storage requirements.

Experimentation with data orderings is often credited to somewhat obscure work carried out by Morton in the mid 1960's for the Canada Geographic Information System (cited in Mark, 1986; Goodchild and Grandfield, 1983; and Lauzon *et al.*, 1985). Morton's order, which was published in an internal report for IBM Canada, allowed cells which are close together in two dimensional space to be placed in similar positions in the linear sequence of the file (see Figure 2c). Further research into data orderings was undertaken by Goodchild and Grandfield (1983) and Abel and Mark (1990). In Goodchild and Grandfield's experiment, four data orders were empirically tested to determine their compression capability. Goodchild and



(a) Row Order (b) Row-prime Order (c) Morton Order

(d) Hilbert Order (e) Column Order (f) Column-prime Order

Figure 2, Spatial data orderings (after Goodchild and Grandfield, 1983)

Grandfield tested row order, row-prime order, Morton order, and Hilbert order (Figures 2a through 2d).

From Figure 2, it would appear that both Hilbert and Morton orders help to preserve the spatial relationships of the two dimensional raster in the translation to a one dimensional sequence, and, as such, longer run-lengths can be expected. In the experiments conducted by Goodchild and Grandfield (1983) boolean images with varying degrees of spatial homogeneity were used to test the compression capability of the various orders. Their results indicated that for images with a high degree of local spatial homogeneity, storage could be reduced by up to 60% using Hilbert order, and 25% using Morton order, as opposed to using row order. For images with little spatial homogeneity their tests resulted in approximately a 5% reduction from Hilbert order and a 5% increase from Morton order, over row order. A comparative analysis, undertaken by Abel and Mark (1990), found row, row prime and Hilbert orderings to be of equivalent performance, and suggested that storage could by reduced by approximately 40% when used in combination with RLE.

4. Application to climate data

Following the work outlined above, experiments were designed to test the relative merits of six different orders for the storage of climate data. The experiments differ from those described above; rather than encoding boolean images, multicoloured images were tested. Raster images corresponding to annual precipitation (total) and annual mean monthly temperature for the North and South Islands of New Zealand, Australia, Queensland, and Bangladesh were used to test the orders. Each record (cell) in the input images contains a four-byte floating point value, recording either its total rainfall or mean annual temperature.

These test images chosen were selected from the databases of IAM currently under development, and therefore the results of the experiment are directly relevant to ongoing research for national-scale IAM development. Summary data pertaining to each the test images can be seen in Table 3.

Dataset	Spatial Resolution	Cells	Background cells
North Island	0.05	22,701	18,085
South Island	0.05	29,141	22,423
Australia	0.2	24,249	12,916
Queensland	0.06	22,185	3,418
Bangladesh	0.05	12,500	7,736

Table 3, The sample images

The record structure for the RLE file consists of a sequence of five byte records; a four byte colour value, followed by one byte recording the run-length. Due to the overhead of storing run-length, it is possible to actually increase the storage requirement for raster images with little or no homogeneity. The decision to allocate one byte to the run-length variable involves a trade-off; images with a low degree of homogeneity will often not reach the upper run-length limit (255), and therefore increasing the size of this variable would increase the size of each record in the file. On the other hand, images with a high degree of homogeneity will often reach the run-length limit, and require an extra record to take the overflow.

The four orders discussed above were tested, as well as two others; column and column prime ordering (Figures 2e and 2f). Column and column prime ordering operate similarly to row and row prime ordering except that the traversal is from top to bottom rather than left to right. Column and column prime orderings were included as it was expected that they should perform best when the climate data has some degree of longitudinal gradient (ie. the values vary less over latitude than they do over longitude).

5. Results

The results from the experiment are illustrated in Table 4. The values in this table represent the percent reduction in the run-length encoded file from its original size. The graphs in Figures 3 and 4 aid interpretation of the results. From these, we can see that the differences in the comparative performance of the orderings are actually very small. Generally, for both climate variables, row and row prime orders provide the best compression, followed by column and column prime, Hilbert, and lastly, Morton order. It is interesting to note that, in most cases, the two dimensional orderings (Morton and Hilbert) are outperformed by the other orders. This is most likely due in part to the fact that the two dimensional orderings are quadrant recursive, and therefore each image needs to be transposed onto a grid with x and y dimensions of 2^n, thus increasing the number of cells that need to be encoded.

The Bangladesh images show the largest range of compression over the six orders. The most effective orderings are row and row prime for temperature, and conversely for precipitation the most effective orderings are column and column prime. This would seem to indicate the flat topography of Bangladesh has less influence effect on the climate and therefore slight latitudinal and longitudinal gradients exist for temperature and rainfall respectively.

Studying the results further, a lack of difference between row and row prime orderings, and column and column prime orderings is apparent. This can be explained by the fact that all the images, except Queensland, have a land mass surrounded by background values (usually a coastal or country boundary), and therefore the colour values in the images are completely surrounded by background values. In effect, this diminishes the differences between the row and column orderings and their prime variants. In the

	Row	Row Prime	Column	Column Prime	Morton	Hilbert
Temperature						
North Island	73.50	73.50	73.47	73.43	71.83	72.34
South Island	70.57	70.57	70.42	70.42	69.35	69.69
Australia	40.95	40.90	40.60	40.60	39.04	39.78
Queensland	-2.67	-2.23	-2.65	-2.47	-4.88	-3.57
Bangladesh	65.40	65.39	59.26	59.26	55.46	62.47
Precipitation						
North Island	73.50	73.50	73.44	73.40	71.80	72.33
South Island	70.51	70.51	70.32	70.32	69.24	69.58
Australia	41.76	41.71	41.38	41.38	39.70	40.62
Queensland	-3.18	-2.67	0.88	1.05	-3.43	-2.27
Bangladesh	51.16	51.15	58.71	58.71	53.61	54.87

Table 4, Percentage compression of individual images

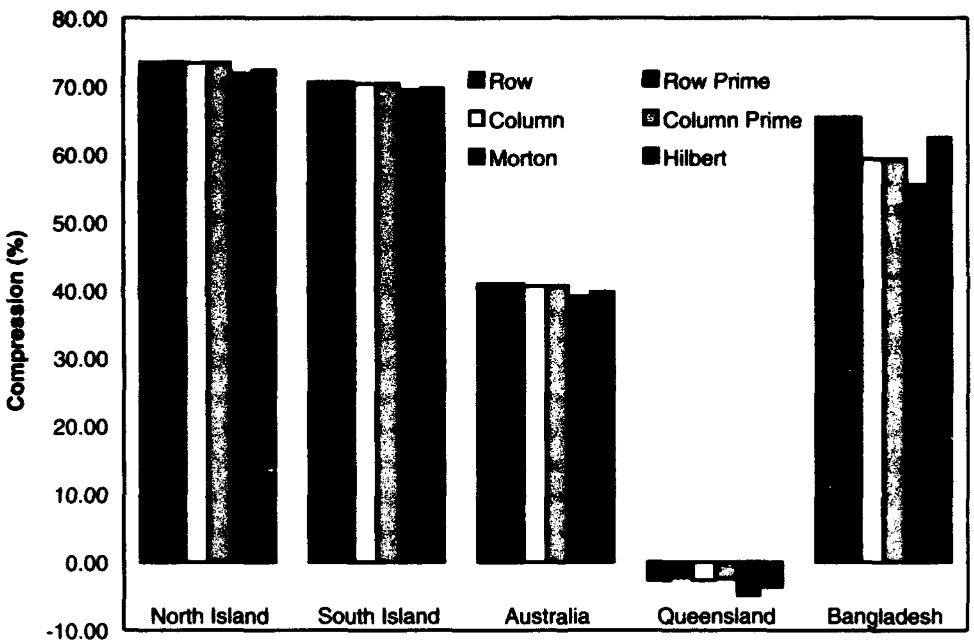


Figure 3, Percentage compression of temperature images

case of the Queensland images, the differences are more pronounced, and for three of the four cases the row and column orderings are less effective than the prime orderings.

In all cases the Queensland images fail to compress to a size smaller than the original file. This could be due to particularly low homogeneity of the colour values for the particular images. However, it is more likely an indication that, in terms of reduction of actual colour values, little is to be gained through any form of data ordering and run-length encoding. The graph in Figure 5 illustrates the relationship between compression (based on the average compression of all the orders) of a given raster and the number

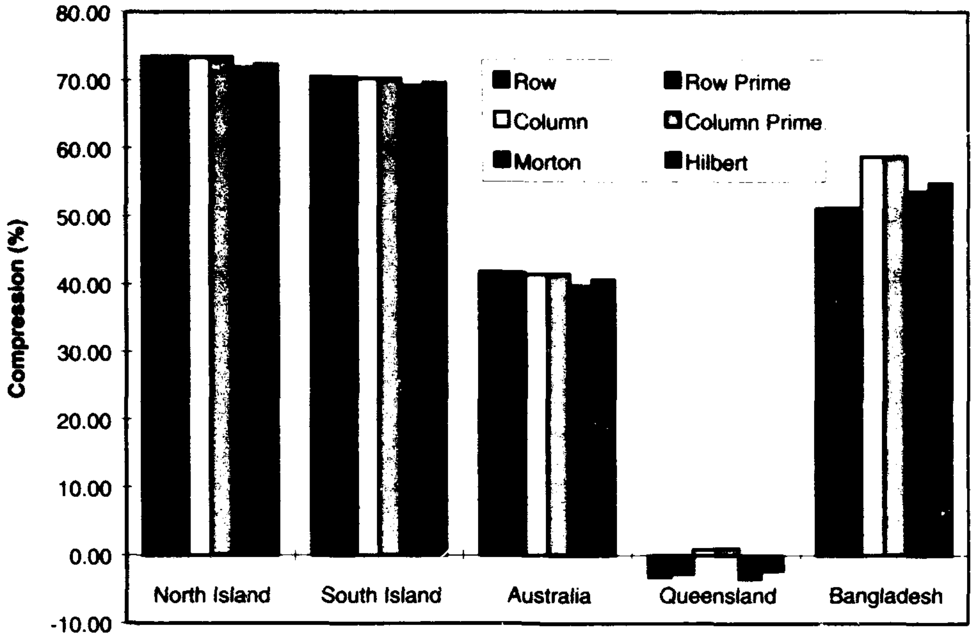


Figure 4. Percentage Compression of Precipitation Images

of background cells. As would be expected, a relationship exists, as background cells are often contiguous and of the same value, therefore large run-lengths should result. However, the relationship is very strong, and in effect, the lack of variation from the fitted line indicates that for the experimental images there is very little spatial homogeneity (and therefore compression) of the coloured cells.

6. Discussion

Climate is driven by solar energy. It is well known that the amount of solar energy received varies latitudinally; the closer one is positioned towards the poles the less energy is received. From this, we could perhaps expect a low degree of latitudinal homogeneity with climate data, and therefore a row order traversal (longitudinal) would result in longer run-lengths. This could perhaps explain why the row and row prime orders marginally outperformed the other tested orders. However, in reality, the problem is not that simple. Climate is highly variable over both latitudinal and longitudinal dimensions, due to the influence of factors such as topography, orography, continentality, and oceans, and this variability undoubtedly influences the poor run-lengths of colour values. Additionally, factors such as spatial resolution of the encoded images will often affect homogeneity — a high resolution image will have a greater degree of spatial auto-correlation than the same image gridded to a coarser resolution.

Another possible contributor to the poor run-lengths for the coloured values lies in the nature of the tested data. Most images of climate variables are interpolated using mathematical procedures from meteorological station weather records. The nature of some of these interpolation algorithms tends to produce images which vary over space between the original site data, sometimes in an unrealistic manner, and fail to adequately represent regional or local scale climates. This is most evident in interpolation algorithms which treat the climate parameter as an independent variable (such as inverse distance weighted algorithms). More advanced techniques, such as co-kriging (Bogaert *et al.*, 1995) or partial thin-plate smoothing splines (Hutchinson, 1995) include the influence of variables, such as elevation, in the interpolation. Often, this type of ap-

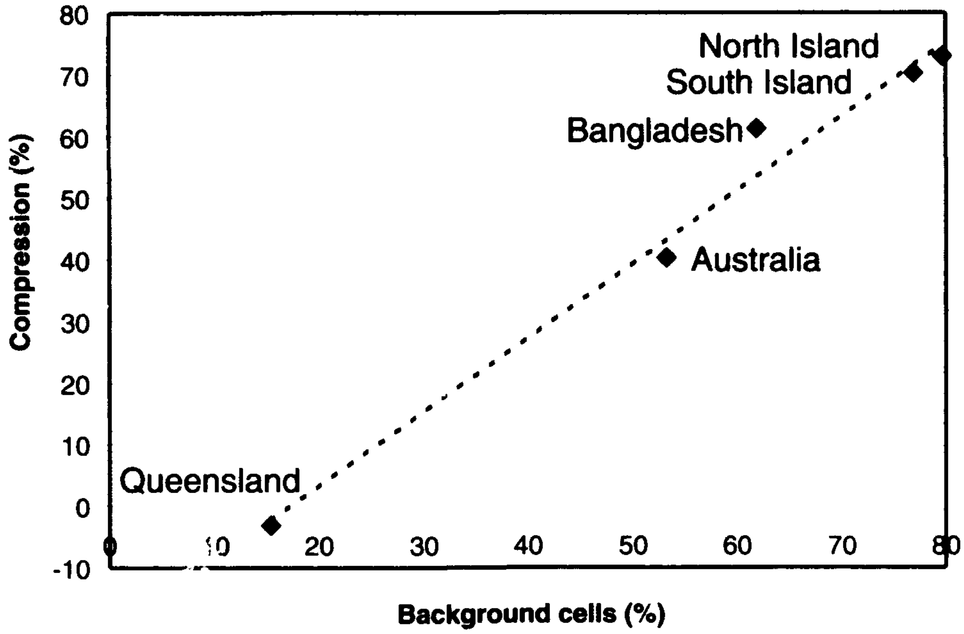


Figure 5. Relationship between compression and number of background cells

proach better captures the spatial variability of climate, and one could expect a higher degree of spatial auto-correlation in the interpolated image. Further to this, interpolated images are only as good as the quality of the point data they are interpolated from. Low density station networks and erroneous site records produce poor quality images with more spatial variability.

7. Conclusions

Overall, the results suggest that, due to the factors outlined in the discussion, it cannot be assumed that any particular data ordering scheme will perform better than any other, or indeed result in any compression at all, with gridded climate data.As such, it would appear that it is not possible to develop a generic algorithm based on one particular data ordering. However, this does not necessarily mean that RLE is an unsuitable technique for application to IAMs. In almost all cases, the size of the original files were substantially reduced through the effective compression of the background values. If it is important that geographic referencing is retained implicitly in the images used by the climate model (for example, if the climate model makes use of images with differing resolutions, projections, or geographic windows), and the images contain a reasonable number of background values, then the use of data orderings and RLE are worthy of consideration.

Acknowledgements

This research was funded by the New Zealand Foundation for Research, Science and Technology (contract C01612).

References

Abel, D.J., Mark, D.M. (1990). A Comparative Analysis of Some Two-Dimensional Orderings. *International Journal of Geographical Information Systems*. 4 (1) pp 21-31

Alcamo, J. (ed.) (1994). *IMAGE 2.0: Integrated Modelling of Global Climate Change*. Kulwer Academic Publishers, Dordrecht.

Bell, S.B.M., Diaz, B.M., Holroyd., F.C. (1988). Cap-

turing Image Syntax Using Tesseral Addressing and Arithmetic. In: Muller J.P. (ed). *Digital Image Processing in Remote Sensing*. Taylor and Francis, London. pp 135-152

Bogaert, P., Mahau, P., Beckers, F. (1995). *The Spatial Interpolation of Agro-Climatic Data: Cokriging Software and Source Code User's Manual.* Agrometeorology Serries Working Paper Number 12. Food and Agriculture Organisation, Rome.

Dowlatabadi, H., Morgan, M.G. (1993). Integrated Assessment of Climate Change. *Science*. (259) pp 1813-1932

Eastman, J.R. (1992a). *IDRISI Users Guide*. Graduate School of Geography, Clark University. Worchester, Massachusetts.

Eastman, J.R. (1992b). *IDRISI Technical Reference*. Graduate School of Geography, Clark University. Worchester, Massachusetts.

Goodchild, M.F., Grandfield, A.W. (1983). Optimising Raster Storage: An Examination of Four Alternatives. *Proceedings, Auto Carto 6*, pp 400-407

Holroyd, F., Bell, S.B.M. (1992). Raster GIS: Models of Raster Encoding. *Computers and Geosciences*. 18 (4) pp 419-426

Hulme, M., Raper, S.C.B., Wigley, T.M.L. (1994). An Integrated Framework to Address Climate Change (ESCAPE) and Further Developments of the Global and Regional Climate Modules (MAGICC). *Energy Policy*. 23 (4/5) pp 347-355

Hutchinson, M.F. (1995). Interpolating Mean Rainfall Using Thin Plate Smoothng Splines. *International Journal of Geographical Information Systems*. 9 (4) pp 385-403

Kenny, G.J., Warrick, R.A., Mitchell, N.D., Mullan, A.B., Salinger, M.J. (1995). CLIMPACTS: An Integrated Model for Assessment of the Effects of Climate Change on the New Zealand Environment. *Journal of Biogeography*. 22 (4/5) pp 883-895

Lauzon, J. P., Mark, D. M., Kikuchi, L., Guevara, J. A. (1985). Two-Dimensional Run-Encoding for Quadtree Representation. *Computer Vision, Graphics, and Image Processing*. (30) pp 56-69

Mark, D.M., Lauzon, J.P. (1984). Linear Quadtrees for Geographic Information Systems. Proceedings, *International Synopsium on Spatial Data Handling*. Vol. 11, Zurich, Switzerland. pp 412-430

Mark, D.M. (1986). The use of Quadtrees in Geographic Information Systems and Spatial Data Handling. *Proceedings, Auto Carto 8*, pp 517-527

Osborn, T.J. and Wigley, T.M.L. (1994). A simple model for estimating methane concentration and lifetime variations. *Climate Dynamics*. (9) pp 181-193

Warrick, R. W., Kenny, G. J., Sims, G.C., Ericksen, N.E., Ahmad, Q.K., Mirza, M.Q. (1996). Integrated Model Systems for National Assessments of the Effects of Climate Change: Applications in New Zealand and Bangladesh. *Journal of Water, Air, and Soil Pollution*. 92 (1/2) pp 215-227

Weyant, J., Davidson, O., Dowlatabadi, H., Edmonds, J., Grubb, M., Parson, E.A., Richels, R., Rotmans, J., Shulka, P.R., Tol, R.S.J., Cline, W., Fankhauser, S. (1996). Integrated Assessment of Climate Change: An Overview and Comparison of Approached and Results. In: Brunc, J.P., Lee, H., Haites, E.F. (eds). Climate Change 1995: Economic and Social Dimensions of Climate Change. Cambridge University Press, Cambridge. pp 364-396

Wigley, T.M.L. and Raper, S.C.B. (1992). Implications for climate and sea level of revised IPCC emissions scenarios. *Nature*. (357) pp 293-300

Wigley, T.M.L. (1993). Balancing the carbon budget. Implications for projections of future carbon dioxide concentration changes. *Tellus*. (45B) pp 409-425

Wigley, T.M.L. and Raper, S.C.B. (1993). Future changes in global-mean temperature and sea level. In: Warrick, R.A., Barrow, E.M. and Wigley, T.M.L. (eds). *Climate and Sea Level Change: Observations, Projections and Implications*. Cambridge University Press, Cambridge. pp 111-133

Interoperable GIS and Spatial Process Modelling

Andrew J. Marr, Richard T. Pascoe, and George L. Benwell
Spatial Information Research Centre,
Department of Computer and Information Science,
University of Otago, PO Box 56, Dunedin, New Zealand,
Tel. +64 (3) 479 8301, Fax. +64 (3) 479 8311
ajmarr, rpascoe@commerce.otago.ac.nz
gbenwell@gandalf.otago.ac.nz

Presented at the second annual conference of GeoComputation '97 & SIRC '97,
University of Otago, New Zealand, 26-29 August 1997

1. Abstract

Recent developments in GIS have focused on the need for technically unrestricted interchange of both spatial data and traditional GIS operations and analysis. In this paper it is asserted that while research in these fields are well advanced, parallel developments in the area of collaborative spatial process modelling development are becoming more reliant on free exchange of both data and models. It is suggested that as these two fields of research advance, the distinction between the two will be blurred. A proposal is put forward for the construction of a system-independent spatial process modelling tool capable of integrating the transfer of data and operations as well as other process modelling functions to complete desired outcomes.

2. Introduction

The development of techniques for access and utilisation of remotely distributed spatial databases via global networks is potentially a 'great leap' for the GIS field (Thoen 1995). The potential benefits of research in this area are the construction of platform independent methods of spatial analysis incorporating the convenient and transparent integration of disparate data sets, and real-time display of queries. In addition to the clear benefits of research in this area, further use of closed proprietary vendor formats is being seen by many organisations as a restrictive practice adding to the desire for open systems (Ayers 1995).

Many current transfer problems stem from the existence of the legacy file formats and complex data transformations required for usability. A similar inefficiency and duplication can be seen in GIS functionality (Astroth 1995). Because of these limitations, Astroth argues that the potential for the further integration of spatial resources has been restricted. While the current developments of the Open GIS Consortium (OGC 1996) in particular would appear promising, the extent of this problem suggests that much more work is required. In this paper, a brief overview of Open GIS will be given followed by the highlighting of some developments in the area of data and process transfer. The paper will then focus on some recent research findings in spatial process modeling before proposing the development of a process modelling tool that attempts combine the two fields.

3. Open GIS: Transfer and Interoperability

The vision of the Open GIS consortium is '...the full integration of geospatial data and geoprocessing resources into mainstream computing and the widespread use of interoperable, commercial geoprocessing software throughout the information infrastructure' (OGC 1996). The proposal is made that GIS software development take the form of 'plug and play' modules leaving the user free to select the best component to solve a specific problem (Glover 1995). The principal thrust behind the Open GIS

	Transfer	Interoperability
Scope	Data, no Process	Data and Process
Data Unit	Dataset	Object (Dataset or Lower)
Communication	Blind (One Way)	Negotiated (Two Way)
Integration	In Target System	In Server or During Communication

Table 1 - Differences Between Transfer and Interoperability from (Glover 1995)

initiative is the development of GIS interoperability (transfer of data and process) rather than just the transfer of straight data. Table 1 details the differences between the transfer of data and interoperability.

'Interoperability' allows for the analysis of data in addition to the straight exchange. The transfer of these two components (data and process) will now be examined with discussion on some of the related issues, and the most promising route for future research.

4. *Transfer of Data*

"Data are the raw facts entered into the computer" (Shore 1988, p10). In GIS terms, data has traditionally been viewed as the 'raw facts' in the structure of fixed proprietary vendor formats. These formats have resulted from the general evolutionary nature of GIS development itself. Because of the 'barriers' (Glover 1995) created by use of different non-interchangeable vendor formats, efforts to overcome these differences have traditionally been time consuming, difficult and resource intensive. While the development of interchange standards such as the Spatial Data Transfer Standard (SDTS, USGS 1996) are useful for bulk transfer, their use is very limited when attempting online transfer (Ayers 1995). This is because the use of a standard '....requires an extra step, can lose data and create inaccuracies, and requires a lengthy import process....' (Ayers 1995, p60).

Recent developments, possibly spurred on by the Open GIS initiative (OGC 1996) have seen some software vendors starting to tackle this problem (Strand 1996). The GeoMedia product launched by Intergraph in March 1997, features limited data access to other vendor formats, through the data warehousing capabilities (Intergraph 1997).

One solution to the current proprietary format exchange problem, is the use one of a growing number of spatial data interchange software package such as FME (1997) or Blue Marble (1997).

> *'The Feature Manipulation Engine (FME) is a sophisticated configurable spatial data processor and translator. The FME facilitates powerful interoperability between diverse systems, and can be used as the backbone of an on-demand mapping system.'* (FME 1997)

If this type of development is a prelude to future initiatives by other solution providers, and more particularly, GIS providers, then this is evidence to suggest that vendor data integration research is progressing favourably. This is consumer driven and reflects a changing attitude towards the importance in sharing data resources (Marr 1996).

5. *Transfer of GIS Operations*

Albrecht (1996, p36) derives a 'conclusive list of Universal GIS Operations' shown in table 2. According to Albrecht, these operations represent the building blocks from which more complex operations may be constructed. These operations were identified by Albrecht from the processes commonly found in existing GIS software and have each been defined algebraically. Algebraic specifications were chosen because they are relatively easy to implement by a functional programming language, and provide unequivocal function definitions.

Once these operations have been defined in this manner it is suggested by the authors that a major step has been made towards the free exchange of GIS operations. Since mathematical definitions exist for GIS operations, a good foundation has been made towards the creation of systems for the remote control of these primary operators on host spatial databases. Alternatively, mechanisms may

Search	Interpolation; Search-by-region; Search-by-attribute; (Re-)Classification
Locational Analysis	Buffer; Corridor; Overlay; Voronoi/Thiessen
Terrain Analysis	Slope/Aspect; Catchment/Basins; Drainage/Network; ViewShed
Distribution/Neighbourhood	Cost/Diffusion/Spread; Proximity; Nearest-Neighbor
Spatial Analysis	Multivariate analysis; Pattern/Dispersion; Centrality/Connectedness; Shape
Measurements	Measurements

Table 2 - Universal GIS Operators from (Albrecht 1996)

be put in place to send locally stored operations to act on the remote data assuming security is not compromised. This is analogous to the use of Java applets, but are too restricted in their operations on the client machine.

6. Spatial Process Modelling

There are many actual and potential applications for spatial process modelling, and as such, research into the construction of generic process modelling tools and methods with maximum useability and flexibility are preferable. Parks (1993) recognised that the majority of recent spatial modelling research has focused on environmental issues. This appears to have resulted in a bias towards environmental modelling development as presented in the literature. It is argued here that much of the work reported has general application and thus no distinction is made.

There is great potential for spatial processing software that integrate the benefits of GIS with the process analysis capabilities of modelling software (Abel *et al.*, 1997; Bennett 1997). Parks (1993) argues that with appropriate planning, modelling and GIS technology may '...cross-fertilize and mutually reinforce each other' (p31) and that both will be made more robust by '...their linkage and coevolution' (p33). According to Abel *et al.* (1997), this integration in the past has been technically difficult to achieve. Abel *et al.* (1997, p5) argue that many examples of GIS and modelling systems integration '...are typically specific to the component subsystems and to the narrow application focus of the integrated system'.

Ball (1994, p346), defines a good model '...as one that is capable of reproducing the observed changes in a natural system, while producing insight into the dynamics of the system'. This implies that the model has two functions. First, to simulate and predict based on observed processes, and second, provide detailed understanding of the inter-relationships among variables and processes described by the model. Simulation modelling must '...describe, explain, and predict the behaviour of the real system' (Hoover *et al.* 1989, p5) and '...requires that the model indicates the passage of time through the change in one or more variables as defined by the process description' (Ball 1994, p347). Ideally, in an integrated geographical modelling system (GMS), as described by Bennett (1997, p337), '...users should be able to visualize ongoing simulations and suspend the simulation process to query intermediate results, investigate key spatial/temporal relations, and even modify the underlying models used to simulate geographical processes'.

The limited development of these models in the past is according to Maxwell *et al.* (1995, p247) due to '...the large amount of input data required, the difficulty of even large mainframe serial computers in dealing with large spatial arrays, and the conceptual complexity involved in writing, debugging, and calibrating very large simulation programs'. An accepted method of reducing program complexity argue Maxwell *et al.* (p251) involves '...structuring the model as set of distinct modules with well-defined interface.'. Maxwell *et al.* suggest that the use of a modular hierarchical approach permits collaborative model research, and simpler design, testing, and implementation. Bennett (1997) and Maxwell *et al.* (1995) advocate the use of modelbase management systems to store, manipulate, and retrieve models. Bennett (p339) states that 'by managing models like data, model redundancy is reduced and model consistency is enhanced'.

Maxwell *et al.* (1996) suggest that one way to develop sim-

pler process model design tools, is to construct suitable graphical interfaces for the display and manipulation of structure and dynamics. Albrecht *et al.* (1997, p158) suggest the use of a '...flow charting environment on top of existing standard GIS that allow the user to develop workflows visually.'. In addition Bennett (1997) and Parks (1993) assert the need for artificial intelligence, expert systems, and agents to guide non-expert users in the appropriate handling of these tools and reduce the need for the writing of complex computer code.

7. Major Issues Spatial Process Modelling to be Resolved

Besides the difficulties in linking GIS functionality to process modelling software discussed in the previous section, there are potential problems in the standardisation of process model description. This is highlighted by Abel *et al.*(1997) who recognises the need for compatibility with legacy models and identifies the requirement in many cases to 're-use' rather than 're-implement'. To promote internationally collaborative development of sophisticated modular process models as supported by Maxwell *et al.* (1995), there needs to consistency. More specifically, if there can be agreement on the format of a modelling language, then unrestricted development of modelling tools may take place. Other areas for further research include the need for transparent access for spatial modelling tool during operation to high performance computers, support for differing spatial representations, and temporal dynamic modes (Maxwell *et al.*, 1996). In addition to these improvements, Bennett (1997) argues the need for developments in four-dimensional data structures, improvements in scientific visualisation, equation generation, and model validation and calibration.

8. Spatial Process Modelling System II

It is proposed to construct a system to design spatial process models, permit sharing of model structure, and execute the process model on user selected data.. The functionally independent components of the system will in the form of services. Services will initially comprise model design, model interpretation, GIS operations, data conversion, and visualisation. These self contained modules would be able to be enhanced and replaced as required without affecting the rest of the modelling system. This modelling system is viewed as a consolidation and extension of the SPMS modelling system (Mann, 1997).

For illustrative purposes, a very simple model representing a standard cartographic modelling process has been shown in Figure 1. The purpose of this model is to select suitable parachute drop sites given specific criteria relating to maximum ground slope, and proximity close to or away from, air corridors, access roads, and waterways. Figure 1 represents a screen shot of a non-functional prototype of a model design and construction interface and one of the services in the modelling system. The current version of this interface is written in Visual Basic, but is currently in the process of being converted to Java for maximum cross-platform portability. The data conversion service will be provided by FME (1997). To enable this, a specific interface will be constructed between this, and the model interpretation service. FME, does have the minor limitation in that it is platform dependent requiring Solaris or Windows95/NT, but it is believed that with the use of self contained services, future versions of the software may remove this limitation.

Using this object based interface, the user is able to place objects from the menu onto the modelling area. The object may take the form of data inputs, data outputs, spatial operators (defined by Albrecht 1996), and mathematical operators. In addition other specialised objects include time constraints and other sub-modelling components. Links are drawn between the objects, but these serve only to indicate the sequence of processing steps which may be forward or reverse (which provides feedback loops). In this example, the four inputs, slope, airspace, road, and hydro, are linked to either a buffer or overlay spatial operation, concluding in the desired output. The required parameters for each object may be specified by clicking on an object. These parameters vary according to the nature of the object.

Once model design has taken place, it is intended that the

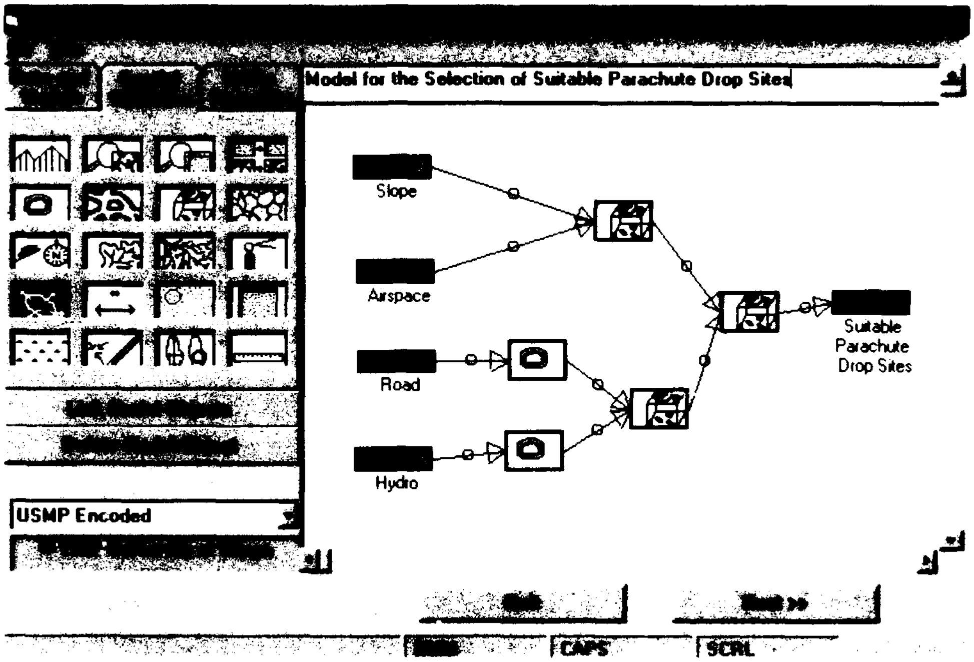


Figure 1 - Prototype Spatial Model Design and Construction Interface

structure may be distributed widely and reused. A potential user may receive the model structure file, and either include it in their own model construction or send it to the implementation service. When opened the implementation service reads the file and determines the required inputs and outputs, before creating a dynamic interface for the specification of required data sources. Figure 2 is a non-functional example of such an interface as it would relate to the previous parachute dropsite problem.

In addition to the specification of data sources and destinations, the interface would also provide detailed model descriptions and limitations, and options for how the processing should proceed. This format will allow users to insert their own data for full utilisation of the model. There are a significant problems to be resolved such as data type specific processing, security, meta-data, and version control.

9. *Conclusion*

The development of highly sophisticated spatial process modelling techniques, involving the modular and distributed amalgamation of GIS and modelling software capabilities is progressing rapidly. At the same time research is continuing into GIS interoperability, representing the unrestricted exchange of data and process.

In this paper the role of spatial model interchange in relation to the transfer of spatial data and operations has been discussed. Analysis of the features of both suggest a blurring of the differences between interoperable GIS and advanced spatial process modelling systems. A potential conceptual strategy has been discussed, that would integrate some of the more recent research and tools, to advance the knowledge in this area. For the success of such a project it is recognised that ongoing work in the interchange of data and operations is paramount.

10. *Acknowledgment*

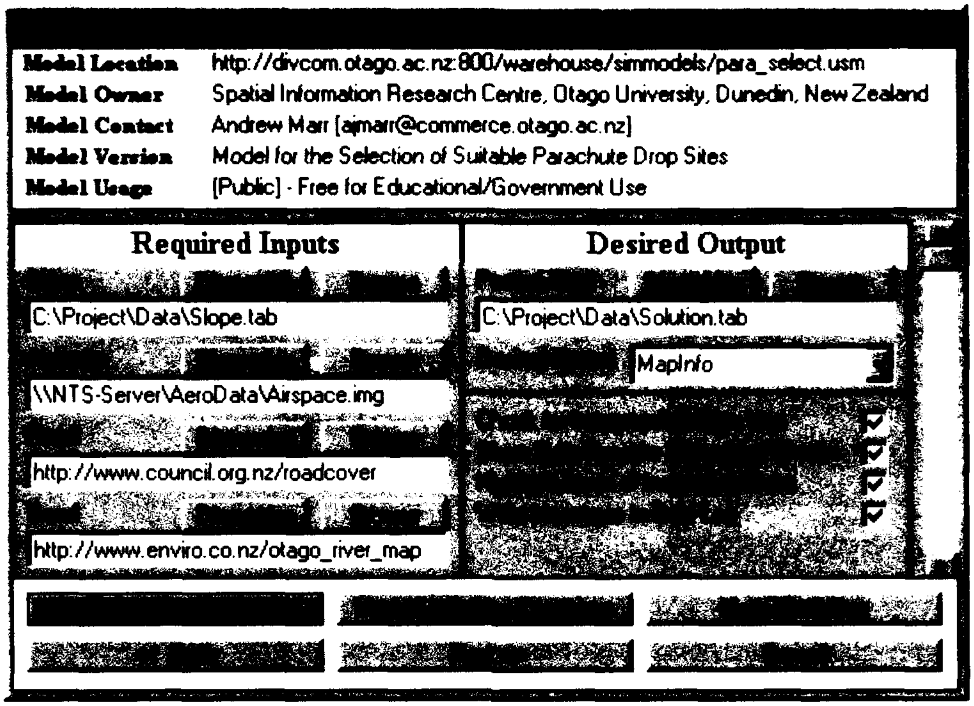


Figure 2 - Prototype Spatial Model Implementation Interface

The authors would like to thank Samuel Mann for providing assistance in the generation of components used in this system based on research conducted for the SPMS system.

11. References

Abel, D., Taylor, K., and Kuo, D. (1997) Integrating Modelling Systems for Environmental Management Information Systems. *SIGMOD - Quarterly Publication of the Association for Computing Machinery Special Interest Group on Management of Data*, 26(1): 5 - 10.

Albrecht, J., Jung, S., and Mann, S. (1997) VGIS: a GIS Shell for the Conceptual Design of Environmental Models, In: *Innovations in GIS 4*, Kemp, Z. ed. London, Taylor & Francis Ltd., p154 - 165.

Albrecht, J. (1996) Universal GIS-Operations - A Task-Oriented Systematization of Data Structure-Independent GIS Functionality Leading Towards a Geographic Modeling Language. University of Osnabrueck, Germany, Unpublished Dissertation Thesis, 99p.

Astroth, J.H. (1995) OGIS: Planting the Seeds for Rapid Market Growth. *Geo Info Systems*, 5(1): 55, 58.

Ayers, L.F. (1995) ViewPoint: What Are Open Geographic Data?. *Geo Info Systems*, 5(1): 60.

Ball, G.L. (1994) Ecosystem Modeling with GIS. *Environmental Management*, 18(3): 345 - 349.

Bennett, D.A. (1997) A Framework for the Integration of Geographical Information Systems and Modelbase Management. *International Journal of Geographical Information Science*, 11(4): 337 - 357.

BlueMarble (1997) *Geographic Translator*. Blue Marble Geographics, *URL= http://www.bluemarblegeo.com/apptrans.htm, geoinfo@bluemarblegeo.com.*

FME (1997) *FME - The Universal Spatial Data Translator*. Safe Software Inc., *URL= http://www.safe.com,*

info@safe.com.

Glover, J. (1995) The Need for Open GIS, Part 1: The Integration Challenge, *Mapping Awareness*, 9(8): 30 -33.

Hoover, S.V. and Perry, R.F. (1989) *Simulation - A Problem-Solving Approach*, Addison-Wesley Publishing Company, Inc., Massachusetts, 400p.

Intergraph (1997) *Intergraph Online - The GeoMedia web site*. Intergraph Corporation, *URL= http://www.intergraph.com/iss/geomedia/, webmaster@intergraph.com.*

Mann, S. (1997) Spatial Process Modelling for Regional Environmental Decision-Making. University of Otago, Unpublished PhD Thesis, 269p.

Marr, A.J. (1996) Geographic Information Systems Maturity in New Zealand Local Government. University of Otago, Unpublished M.Sc. Thesis, 126p, *URL= http://divcom.otago.ac.nz:800/sirc/webpages/people/andrew/papers/masters/masters.pdf, ajmarr@commerce.otago.ac.nz.*

Maxwell, T. and Costanza, R. (1996) *Facilitating High Performance, Collaborative Spatial Modeling*. 3rd International Conference on Integrating Geographic Information Systems and Environmental Modelling, *URL= http://www.ncgia.ucsb.edu/conf/SANTA_FE_CD-ROM/sf_papers/maxwell_tom/ecohpc.html, mailto:maxwell@kabir.cbl.cees.edu.*

Maxwell, T. and Costanza, R. (1995) Distributed Modular Spatial Ecosystem Modeling, *International Journal of Computer Simulation*, 5(3): 247 - 262.

OGC (1996) *OpenGIS Overview*. Open GIS Consortium, Inc., *URL= http://www.opengis.org/overview.html, gbuehler@mail.opengis.org.*

Parks, B. (1993) The Need for Integration, In: *Environmental Modeling with GIS*, Goodchild, M., Parks, B. and Steyaert, L. eds, New York, Oxford University Press, p31 - 34.

Shore, B. (1988) *Introduction to Computer Information Systems*, Holt, Rinehart and Winston, Inc., New York, p540.

Strand, E.J. (1996) Open GIS Client/Server Products Remain Elusive, *GIS World*, 9(3): 36 - 38.

Thoen, B. (1995) Interactive Mapping and GIS Thrive on the Web, *Gis world*, 8(10): 58-59.

USGS (1996) *SDTS Homepage*. U.S. Geological Survey, *URL= http://mcmcweb.er.usgs.gov/sdts/, sdts@usgs.gov.*

A Method for the Integration of Existing GIS and Modelling Systems.

Paul Yates and Ian Bishop
Department of Geomatics,
University of Melbourne,
Parkville, Victoria, Australia, 3052.
pmy@sli.unimelb.edu.au
I.Bishop@engineering.unimelb.edu.au

Presented at the second annual conference of GeoComputation '97 & SIRC '97, University of Otago, New Zealand, 26-29 August 1997

Abstract

The integration of existing modelling and geographic information systems (including GIS) is an important activity that enhances the value of these systems. In this paper, we present a simple and comprehensive approach for the integration of separately developed software systems. Any information system can be integrated using our methodology without the complexities introduced by providing an interpretation of a universal language. The design of our integration methodology consists of four separate components, the protocol for communication, a message queuing system, wrapping software and an integration manager. Relevant conceptual models and implementation techniques are discussed in the paper. We also describe some examples of the software we have successfully integrated and present an example script for managing a simple integration activity.

1 Introduction

Currently, there are many spatial database management systems (GIS), aspatial database management systems and modelling systems used in modelling activities. In general, these software systems are, and have been, developed independently with their own specifications, interfaces, data models and data types. As it is often the case that information and operations needed for a particular task exist in separate information systems, there is a need to integrate these information systems. We note that there are many different meanings given to the term integration. These range from providing a simple methodology to allow the cohesive operation of the separate information systems, through to providing a high level language and data model that encompasses all the operations and complexities of each separate information system (e.g., the "Universal GIS Operators" described by Albrecht (1995) and the Open Geodata Model of the Open GIS Consortium (1996)).

An example of the need for integration is found in the urban modelling area. Wegener (1994) surveys the state of the art in operational urban models. In the survey several different urban subsystems are identified and several urban modelling systems are identified that model some or all of the urban subsystems. There is a strong interest in being able to integrate these models with GIS. Also, where models do not cover all urban subsystems, there is a need to be able to integrate the models so as to increase the number of subsystems covered.

In this paper, we present a simple and comprehensive approach for the integration of separately developed software systems. Any information system can be integrated using our methodology without the complexities introduced by providing an interpretation of a universal language.

An example of such complexity is the need to translate the universal language into a language understood by an individual information system. In the research area of federated information systems based on relational and object oriented databases the translation process is less difficult

as there already exist standard languages such as SQL and OQL understood by most, if not all, databases that need to be integrated. Given this fact, during the integration process the developer can concentrate on other problems related more specifically to the data and data structure (such as schema translation and schema integration (Sheth & Larson 1990)). When integrating software supporting geographic information systems and modelling there is seldom a common language to the systems being integrated. This means that the developer must be concerned with the language, data model and data structures of each system. This is a primary difference from the standard requirements for the integration of relational or object oriented database systems.

The OGIS guide (Open GIS Consortium 1996) identifies several software layers in the design of integration software. These include, the presentation layer, the application and application server layer, the spatial data access provider layer, the database layer and the hardware and network layer. Our proposed methodology concentrates on the implementation of the application and database layer. We do not directly address the important issues relating to the development of a high level language and data model of the spatial data access provider layer. It is significant to note that to be able to integrate information systems at a high level, there is also a need to integrate to the level described in this paper. What we directly address in this paper are issues relating to the design and implementation of a system that allows concurrent access to data and programs (in their current form) provided in disparate information systems. Most importantly, we use languages native to each of the individual information systems to access the data and programs.

There are many important issues to consider in the development of a system to integrate disparate information systems. These issues include, data access, interoperability, integration process management, user interface design and security. Data access encompasses the requirements for transformation between data models and translation between data types as well as the communication of the data between software systems. By interoperability we mean the ability to access the modelling systems (programs) available in different information systems. It is important that there exists the ability to control the integration process, i.e., there is a need to provide for the specification of the steps required to perform a given task needing integration of separate software systems. Currently, we are not immediately concerned with the provision of a user interface; at present we provide an interpreter for a small language to manage the integration process. Eventually, we intend to provide a "drag-and-drop" type interface linking models to data sets. However, to provide such an interface we will need to address issues related to providing a universal language for integration. Another issue that is beyond the scope of our current implementation is security. This includes concepts relating to the rights to use the data and software, and auditing of such use.

2 Background

2.1 Conceptual Models

Integration of existing software systems has been the subject of much recent research. In particular, Abel, Taylor & Kuo (1997) develop a theory for the integration of modelling systems for environmental management information systems. In the model several generic concepts relating to software integration are identified. The concepts of an object, problem, solved problem, solver, execution plan and a well-defined problem are all defined. A significant point is that they equate the concept of a problem with the concept of a query in a database system. A solver then provides a solution to a problem in a similar fashion to the way a database provides an answer to a query. Their model provides a conceptual framework that enables both a proper description of a given integration activity as well as a description of the software components used to develop a solution for an integration problem.

Wiederhold (1992) develops the concept of mediators for information systems. The paper discusses an architecture for an information system consisting of three layers, a user layer, a mediator layer and a base layer (possibly consisting of multiple databases). The mediator layer of an information system sits between a user layer and a base layer. It is

the responsibility of the mediator layer to accept requests, distribute the requests to the appropriate information system in the base layer and collate and return the result of a request to the user layer. The mediator layer may make use of knowledge about the request (and the data required to answer the request) to decide how to distribute the request to the base layer. Buneman, Raschid & Ullman (1997) propose a "mediator language" in which it is possible to describe the data structures and data models that are part of a given information system. In addition, the language allows for the expression of database queries which can then be passed to a given information system.

Making an independently developed software system communicate often requires the system to be "wrapped", i.e., a piece of software is developed that communicates to external processes as well as controlling the systems being integrated. The use of wrappers is a common component in a number of software architectures used for integrating software (Buneman et al. 1997, Ishikawa, Furudate & Uemura 1997, Papakonstantinou, Gupta, Garcia-Molina & Ullman 1995, Wiederhold 1992). The wrapping software provides a shell around the software to be integrated, providing a point of access to the integrated software.

2.2 Implementation

Implementation of a system for the integrating software inherently makes use of pre-existing approaches that support the development of distributed systems software. Examples of such approaches include the Remote Procedure Call (RPC) (Comer & Stevens 1993), the message queuing model (Blakeley, Harris & Lewis 1995).

In the remote procedure call paradigm, calls to procedures that do not exist in the calling program are passed to a remote program for execution. The calling program (or client program) then waits for the completion of the called procedure before continuing execution (exactly as it would if the procedure was local). Data is passed to and from the remote program using a common data representation (such as the External Data Representation XDR). Such a method of communication is called synchronous, as the calling program waits until the called procedure returns before continuing. In a message queuing system there is a message queue associated with each participating system. Using message queuing, a system communicates by placing messages on a queue associated with the system with which it needs to communicate. The called system will then retrieve the message from its queue when it is ready. The calling system is free to continue processing or wait depending on its own needs. Hence, the communication can be synchronous (as with the RPC mechanism) or asynchronous. Commonly, the calling program only directly communicates with a program, called the queue manager, whose specific duty is to manage the queues associated with each program.

Blakeley (Blakeley et al. 1995) defines a criteria for the selection of the appropriate style of communication. Significantly, the use of message queuing systems is most appropriate when there is a mixture of application types, old and new programs and network types and where the programs are highly independent. These criteria are common with the requirements for integrating GIS and modelling systems.

Once a message queuing communication process has been established, each integrated process must be capable of understanding the message it is passed. The usual method of passing understandable messages is to define a protocol. The definition of a protocol defines the structure and interpretation of the messages that can be passed amongst the integrated systems. For example, suppose we have a protocol containing a command "exec" with one string parameter that asks for the execution of the string it is passed. A message "exec union covera coverb coverab" passed to an ARC/INFO process would instruct ARC/INFO to execute a union operation between the coverages a and b.

3 Architecture

Several conceptual models and methods for implementation were identified in the previous section. Issues that we considered important while designing and implementing the integration software are:

(1) the integration of new software components should require the minimum amount of programming effort;
(2) the protocol for communication should contain the smallest set number of commands necessary to enable integration (i.e., the smallest set enabling one software component to make calls to functions provided in another software); and
(3) there should exist some ability to control the integration process through the use of an integration language.

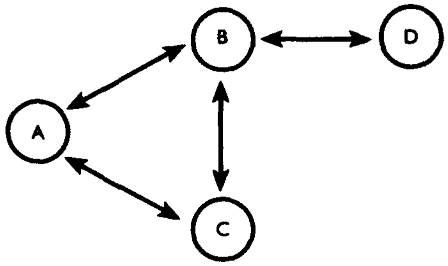


Figure 1: Asynchronous messaging between process.

3.1 Components

The design of our integration methodology consists of four separate components, the protocol for communication, a message queuing system, wrapping software and an integration manager. We begin with the definition of a protocol for which the separate systems communicate. To meet the minimum requirements, we have developed a protocol that includes methods to: establish and close communication; and execute programs and/or scripts. The parameters of the establish communication command include the specific location queue manager and the location of the process that is making itself available for integration. The contents of the execute command is text understood by the integrated process. To accept and request data we make use of the existing file transfer protocol definition.

The second component consists of a message queuing system. The message queuing system manages a queue for each path of communication that is established. Messages are passed to the message queuing system from an integrated system and placed on the appropriate queue. They then remain on the queue until they are retrieved by the appropriate system. Hence, the communication between integrated systems is asynchronous. It is important to note that we specify that the communication between a given system to be integrated and the message queuing system is synchronous. Figure 1 shows an example of the communication between four systems. The communication at this level is asynchronous. Figure 2 shows the actual communication paths that exist for the abstract communication paths shown in figure 1. Messages from A to B are placed on B's queue. Messages from B to A are placed on A's queue.

The third component consists of software to implement the wrapping of the information systems to integrate. At the moment this software is set up to communicate with the queue manager. We do not specify the method of communication between the wrapping software and the system to integrate as this depends upon the system being integrated.

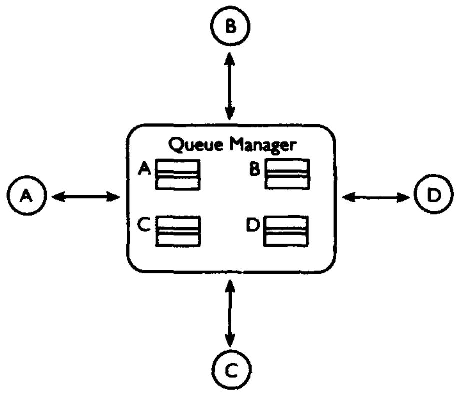


Figure 2: Synchronous messaging between processes and the queue manager.

For example, ArcView has the ability to communicate using the RPC mechanism, other software may not have this ability or there may be some other preferred method of communication. The wrapping software does not interpret the string passed for execution. That string is passed to the wrapped software for interpretation.

The fourth component of our design consists of the specification of a small interpreted language to manage the integration process. The basic elements of the language are

the commands, the open, exec, send and mode. The use of the open command allows for the definition of multiple connections to different software systems. Messages can be sent to defined connections through the exec command. Such messages sent through the connection are specific commands relevant to the particular software system on the connection, hence, the use of "exec" for the name of the command. For example, for the connection to an ARC/INFO session a message may contain a particular ARC/INFO command or macro. As several open connections to different software systems can be defined, a single script written in the integration language can contain a mix of the languages available to different software systems. The mode commands provides a method to specify the type of communication (synchronous or asynchronous). This command is useful for the cases in which the processing of a given script must be done synchronously. Finally, the language also includes commands to send data to a connection and request data from a connection. Figure 3 shows a simple script for running an urban model on data

```
mode.sync     # Set to synchronous mode of communication mode.sync
# Declare connection to arc/info on machine scamper and
# an urban modelling package on machine daisy and buttercup
open arc scamper, um1 daisy, um2 buttercup
# Export data in preparation for use in urban model
arc.exec workspace /usr9/pmy/urban/in ; gridascii house house.asc ;
gridascii pop pop.asc ; gridascii emp emp.asc
# Send data from scamper to daisy and buttercup
send scamper:/usr9/pmy/urban/in/*.asc daisy:/usr4/people/pmy/umdata/in
send scamper:/usr9/pmy/urban/in/*.asc buttercup:/home/pmy/umdata/in
# Mode can now be asynchronous for the execution of models
mode.async
# Step the urban model by one step ... input parameter 0.05
um1.exec step 1 0.05 house pop emp
# Step the urban model by one step ... input parameter 0.95
um2.exec step 1 0.95 house pop emp
# Reset mode to synchronous (ie wait for models to finish)
mode.sync
# Return stepped data
send daisy:/usr4/people/pmy/umdata/out/*.asc scamper:/usr9/pmy/urban/out1
send buttercup:/home/pmy/umdata/out/*.asc scamper:/usr9/pmy/urban/out2
# Re-import data
arc.exec workspace ../out1 ; asciigrid house.asc house ;
asciigrid pop.asc pop ; asciigrid emp.asc emp
arc.exec workspace ../out2 ; asciigrid house.asc house ;
asciigrid pop.asc pop ; asciigrid emp.asc emp
# Compute the difference in the population values and gridshade it
arc.exec workspace ../out ; grid ; popdiff = ../out1/population - ../out2/population " quit
# Close connections
close um,arc
```

Figure 3: A simple example script for integrating ARC/INFO with an urban modelling package

stored in ARC/INFO. The model is run twice with a different input parameter on different machines. Note the two executions of the model are done simultaneously on different machines.

3.2 Implementation Process

Given the components described in the previous section, the process of integration consists of two steps. First, the wrappers understanding the above protocol are developed for each the information systems. Wrappers can be implemented in a language such as C or using a portable scripting language such as TCL while making use of tools such as Expect for automating interactive applications (Libes 1994). The second step is to write a script for controlling the integration process. This script is interpreted using the integration manager.

Use of the integration manager is not always mandatory. For example, it may be the case that the software being integrated can communicate with the queue manager without the need to use the integration manager. For example, ArcView includes RPC classes and the wrapper understanding the integration protocol can be built using Avenue. Other Avenue scripts may also also place messages on the queue of another process independently of the integration manger. In the case where there does not exist some communication ability within the software being integrated, it may be necessary to drive the process using the integration manager. Consider integrating several executable programs that cannot themselves place messages on a queue. Such examples can be found in urban modelling where different previously developed software may be concerned with modelling different urban subsystems (such as transport and employment). The protocol we use does not have the ability to initiate and control a series of steps executed by separate software systems. To do so, we have introduced the integration manager language. Scripts written in the integration manager language provide the necessary steps to perform a given task.

3.3 Examples

Currently, we have implemented such software to wrap ARC/INFO so as to enable it to be integrated with modelling software written in the unix environment. ARC/INFO, through its inter-application communication (IAC) provides RPC connection to other processes. The wrapper simply interprets the messages passed to it in the following way: a request for a connection starts ARC/INFO in a server mode; an execute command passes the string to the ARC/INFO session; a close connection request closes the ARC/INFO session. Some other software systems that we are looking to integrate includes GENAMAP and Illustra (an object-relational database management system).

Another example can be found in our integration of ArcView (running on a Sun) and a piece of visualisation software developed using the Performer toolkit (on an SGI). This integration only makes use of the queue management software. Both ArcView and the Performer based program act as clients to the queue manager. The integration of these programs was done for a specific project in which a polygon that is selected using ArcView is highlighted in a three-dimensional scene viewed using the Performer toolkit.

4 Summary and Future Directions

In this paper we presented a method to integrate existing information systems. We did not require the implementation of a high level language for integration, but took a more simplistic view of integration concentrating on minimum requirements necessary to enable communication and sharing of procedures between systems. In our design and implementation we have concentrated on issues relating to interoperability. The method is comprehensive in the sense that most (if not all) software can be wrapped with software understanding the protocol we have defined.

Other work on software integration for GIS and modelling systems has concentrated on the definition of a high level language, data models and data structures for integration. We have not addressed specific issues relating to such languages. However, note that the high level languages would need to be translated into calls on the individual systems. This may be possible through the translation of the language into the integration management language

described in this paper. This translated script could then be interpreted using the software describe here.

Although, we have currently implemented both the interpreter for the wrappers and management language in C, there is no reason that we cannot use another language to implement these systems. In fact, we intend to implement the interpreter as a Java applet enabling its use through any system containing a Java interpreter (e.g. Netscape).

References

Abel, D. J., Taylor, K. & Kuo, D. (1997), 'Integrating modelling systems for environmental management information systems', SIGMOD Record 26(1), 5—10.

Albrecht, J. (1995), Universal Analytical GIS Operations: a task-oriented systematisation of data structure-independent GIS functionality leading towards a geographic modelling language., PhD thesis, University of Vechta, Germany.

Blakeley, B., Harris, H. & Lewis, R. (1995), Messaging and queuing using the MQI : concepts & analysis, design & development, McGraw & Hill Book Company Inc., New York.

Buneman, P., Raschid, L. & Ullman, J. (1997), 'Mediator languages — a proposal for a standard', SIGMOD Record.

Comer, D. E. & Stevens, D. L. (1993), Internetworking with TCP/IP Vol III: Client-server programming and applications, Prentice-Hall Inc., New Jersey.

Ishikawa, Y., Furudate, T. & Uemura, S. (1997), A wrapping architecture for IR systems to mediate external structured document sources, in R. Topor & K. Tanaka, eds, 'Database systems for advanced applications '97', Vol. 6 of Advanced Database Research and Development Series, World Scientific Publishing Co., Singapore, pp. 431—440.

Libes, D. (1994), Exploring Expect, O'Rielly & Associates Inc.

Open GIS Consortium (1996), The OpenGIS Guide, Introduction to Interoperable Geoprocessing, Part I., http://www.ogis.org/guide/guide1.htm.

Papakonstantinou, Y., Gupta, A., Garcia-Molina, H. & Ullman, J. (1995), 'A query translation scheme for rapid implementation of wrappers', Lecture Notes in Computer Science 1013, 161—186.

Sheth, A. & Larson, J. (1990), 'Federated database systems for managing distributed, heterogeneous and autonomous databases', ACM Computing Surveys 22(3), 258—274.

Wegener, M. (1994), 'Operational urban models, state of the art', Journal of the American Planning Association 60(1), 17—27.

Wiederhold, G. (1992), 'Mediators in the architecture of furture information systems', IEEE Computer pp. 38—49.

Harnessing Spatial Analysis with GIS to Improve Interpretation of Airborne Geophysical Data

Ann-Marie Anderson-Mayes
Department of Geographical Sciences and Planning,
University of Queensland

Presented at the second annual conference of GeoComputation '97 & SIRC '97, University of Otago, New Zealand, 26-29 August 1997

1. Abstract

Modern airborne geophysical surveys are collecting large quantities of high quality data for applications ranging from mineral exploration to environmental problem solving. As a result, there is a growing need for new interpretation methodologies to maximise the amount of information which can be extracted from survey data. This is especially true in the relatively new environmental application areas where interpretation methodologies are not yet well established.

This paper reports on a research project in which spatial analysis with GIS has been adopted as an approach to improve the interpretation of airborne geophysical data for salinity studies The paper discusses the general and particular interpretation problems for this application; proposes a new methodology for interpretation based on spatial analysis with GIS to address these problems; and concludes with the implications of this work for interpretation of airborne geophysical data for other applications.

Geophysics has long been a field which has made use of leading edge computer-based technology to acquire and process data. However, many areas of the analysis and interpretation of the data are still relying largely on visual interpretation. The advances being in made in computational geography, especially in terms of developing spatial analysis tools on a GIS platform, have the potential to make a significant impact on the interpretation of geophysical data.

2. Introduction

Airborne geophysical data has traditionally been collected for mineral and petroleum exploration studies, but in recent years environmental applications have emerged as an important new application area for this technology. These new applications have in turn driven developments for more sophisticated data acquisition technology. This has been greatly facilitated by the rapid improvement in computer technology over recent decades.

Geophysical data acquisition and data processing have long been fields which have made use of leading edge computer-based technology, however much of the interpretation of the data still relies largely on visual interpretation skills, albeit with the aid of sophisticated digital image processing. In at least one of the new environmental application areas, that of salinity studies, the interpretation is proving difficult to complete using the traditional approach. The aim of the interpretation is to build a picture of the hydrogeological mechanisms contributing to salt degradation at both catchment and paddock scales, and consequently to develop land management plans which address both existing and possible future salt hazard sites. However, the sheer quantity of data to be examined and interpreted presents a significant challenge. A new approach to interpretation is required to meet this challenge and enable effective and efficient extraction of information from the large multivariate geophysical surveys which are typically collected for these studies.

This paper reports on a research project in which spatial analysis with GIS has been adopted as an approach to improve the interpretation of airborne geophysical data for salinity studies. The background to using ge[illegible]ysics for salinity studies will be introduced, and then t[illegible]pretation problems which have arisen with multi[illegible]borne geophysical data sets will be discussed. In order to address these problems, a new interpretation methodology based on spatial analysis with GIS will be proposed. Finally, the implications of this work for interpretation of airborne geophysical data for other applications will be discussed.

3. *Salinity and Geophysics*

In recent years, salt degradation of Australia's land and water resources has been widely recognised as a significant environmental problem, although the causes of human-induced (secondary) salinisation can be traced back to widespread clearing of native vegetation post European settlement. Wood (1924) was one of the first researchers to report observations of a link between clearing of native vegetation and land and stream salinisation in the formal scientific literature. Wood (1924) stated that over the *30 year* period prior to publication of his paper he had observed several instances of land and stream salinisation developing after adjacent tracts of land had been cleared.

Since those early observations, a vast body of research has given us a much better understanding of the causes of salinisation. Secondary salinisation can be classified into two general types depending on the absence or presence of a groundwater system (Williamson, 1990). The former type occurs where over-grazing causes erosion and the saline or sodic subsoils are exposed. The latter type can occur under both irrigated and non-irrigated farming regimes, but in both cases groundwater is a key element for the development and maintenance of salinity. Changes to the hydrologic equilibrium cause increased recharge to the groundwater system and this in turn leads to a rising watertable which remobilises salts stored in the sub-surface. The saline groundwater is discharged via seeps and streams, or evaporation occurs in the areas where the watertable is very shallow (within 2m of the surface). The result is an increased concentration of salt in either streams or soils, causing degradation of water resources and productive agricultural land.

It is this groundwater driven salinisation which is of interest to this project. In recent decades, a number of researchers have demonstrated the effectiveness of ground geophysics in investigations of this type of salinity. For example, Engel *et al.* (1987b) used geophysics to define recharge and discharge areas associated with dryland salinity in the south-west of Western Australia.

In the late 1980s a group of researchers recognised that the scale of the salinity problem in Australia (and worldwide) could not be effectively addressed by high cost, low areal coverage, ground based studies (Street and Roberts, 1994). Airborne geophysical surveys, whilst sacrificing some of the detail of ground based surveys, could provide regional coverage for comparatively low cost and highlight those areas that required more detailed follow-up on the ground.

The work o[illegible] 1987a, 1987b), Street and Engel (1990) and othe[illegible]monstrated that magnetic and electromagnetic measurements provided valuable information about constrictions to groundwater movement and salt storage respectively. Airborne geophysics of this type has traditionally been applied to mineral exploration. Whilst the airborne magnetic technology was immediately applicable to salinity investigations, the same was not true of airborne electromagnetic measurements. The airborne electromagnetic systems in use in Australia in the late 1980s had been designed to probe deep into the earth in search of mineralisation targets such as conductive sulphides. In particular, they had been designed to mask near surface conductivity variations - the very information which is most important in salinity studies.

A collaborative research project (World Geoscience Corporation, CSIRO Division of Exploration Geoscience, CSIRO Division of Water Resources) was established to address this problem. The purpose of the project was to develop a new airborne electromagnetic system, SALTMAP, designed specifically to make high resolution measurements

of the electrical conductivity distribution within the regolith. The principal objective was to assist land care specialists to manage planning and implementation of rehabilitation and protection programs in salt-affected areas, by providing cost-effective and accurate information about salt storage and salt hazards at the catchment scale (Street and Roberts, 1994). Development of the SALTMAP system is now complete and further details can be found in Duncan et al. (1992) and Roberts et al. (1992).

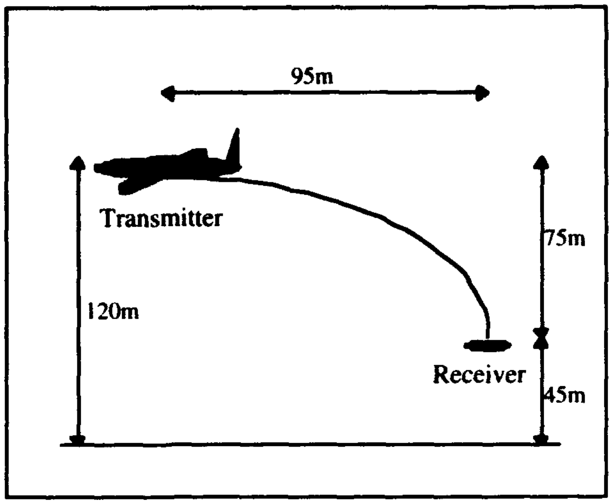


Figure 1 SALTMAP System Geometry (adapted from Roberts et al (1992)).

3.1 A Typical Airborne Geophysical Survey for Salinity

A typical geophysical survey for salinity will include measurement of three geophysical data sets: electromagnetics, magnetics and radiometrics. The data is collected along parallel survey lines, typically spaced 100 or 200 metres apart with measurements recorded along the line every 10 to 15 metres. Depending on the system being flown, nominal flying altitude is between 60 and 120 metres and survey lines are oriented roughly perpendicular to the strike of the general geology, thus maximising the information content of the data sets. For use, this survey line data (known as located data) is usually transformed to raster format (referred to as grids). In addition to the geophysical data, any available surface information relevant to the study can be collected from the relevant government authority, the local landcare organisation(s), and the local farmers.

Electromagnetic measurements are made by the SALTMAP system mounted on a Britten-Norman Trislander aircraft. The approximate flying configuration is shown in Figure 1. SALTMAP is an active measurement system in which a power source connected to a coil mounted on the aircraft structure forms the transmitter, and three perpendicular coils (X, Y, and Z) mounted in a towed bird comprise the receiver. Technical details of the system are reported in Duncan et al. (1992) and Roberts et al. (1992). It is sufficient to note here that a measurement consisting of 100 channels per receiver coil (a total of 300 channels) is recorded every millisecond. However, only the X and Z channels are currently used and, depending on the data, the channels can be binned to a more manageable number (perhaps 15 or 20) or only a selected number of channels are retained for analysis.

Electromagnetic measurements respond to changes in the electrical conductivity of the sub-surface. In most landscapes, the mostly highly conductive material is salt (the only exception to this is some highly conductive clays) (McNeill, 1980) and so where salinisation is a problem, it can usually be assumed that the strongest conductors in the landscape are due to salt. For salinisation to occur there must be a source of salt, so this data is used to map the spatial location and extent of salt storage in the landscape.

Magnetic and radiometric data are collected simultaneously on a single aircraft. The system flies at a nominal height of 70m above ground level. Magnetic measurements are made by a cesium vapour magnetometer installed on a rigid boom at the rear of the aircraft and radiometric measurements are made by a gamma-ray spectrometer installed inside the aircraft.

Magnetic measurements respond to subtle changes in the earth's magnetic field caused by the influence of rocks in the sub-surface on the local magnetic field. The magnetic data can be interpreted (in conjunction with the known geology) to produce an interpreted geology map of the survey area. For salinisation to occur there must be a source of salt and a source of water, in this case groundwater. The geology map, in conjunction with a digital elevation model helps to define the likely groundwater flow for the survey area. In particular, magnetics can identify groundwater barriers. These barriers force groundwater to the surface, and if the groundwater is saline an area of salt degradation results.

Radiometric data, unlike magnetics and electromagnetics, measures only surface phenomena. The energy of gamma rays from decaying radioactive elements is measured and thus a relative distribution of these elements can be mapped. The typical channels of radiometrics which are used are potassium, thorium, uranium, and total count. This information can be used to assist in producing an interpreted soils map (Gourlay, 1996) or to characterise the regolith cover. Such information can be used in conjunction with the electromagnetic data to deduce the potential mobility of the salt. It can also be used to better understand the history of the landscape which can have implications for the potential for salinisation.

It is clear then that the final multivariate geophysical data set comprises perhaps several tens of grids, as well as the digital elevation model. In addition, relevant surface data might include stream network, cadastral data, soils map, regolith map, geology map, existing salt degradation, vegetation cover, and waterlogging. With such a large number of data sets, the interpretation becomes unwieldy and extremely time consuming. Also, a significant risk exists that valuable information, particularly relationships between data sets, might be missed. In the following section, the traditional approach to interpretation is discussed and these interpretation problems are examined in greater detail.

4. Interpretation of Airborne Geophysical Data

In its broadest sense, interpretation can be understood to mean the process of transforming the airborne geophysical data into information. However, this is a long and complicated process and masks the various stages which occur in this transformation. For the purpose of this paper, interpretation will mean the process by which meaning is extracted from one or more final data sets. A final data set will be defined as one which has resulted from passing the raw data through a succession of analyses to

1 remove systematic noise and correct for data acquisition errors (eg. varying altitude);
2 present the data in a useful format (eg. transform line data into gridded data); and,
3 present the data as a useful measurement (eg. electromagnetic data might be transformed to conductivity data).

For a typical geophysical survey for salinity studies, these final data sets will be magnetics and radiometrics grids, and electromagnetics transformed to a suite of conductivity grids. The digital elevation model will also be a grid and the surface data sets will be available either in map or digital form depending on the data source.

The aims of the interpretation are

1 to identify the hydrogeological causes of salinisation in the survey area;
2 to predict and rank all sites at risk of salt degradation based on the hydrogeological interpretation; and,
3 to develop a land management plan based on these results.

The first two interpretation tasks are the focus of this paper as they involve direct interpretation of the airborne geophysical data. For the first, each data set is examined in turn, and from it information relevant to the hydrogeology of the area is extracted. In the case of magnetics, this will involve a full geological interpretation based on the magnetic data, the known geology, and the interpreters own experience and knowledge of the area (Isles *et al.*, 1994). On the other hand, only the areas of high conductivity

(delineating interpreted salt storage) and low conductivity (delineating areas of potential recharge) might be of interest in the electromagnetic data. Alongside these individual interpretations, interpretation of integrated data sets will seek relationships between the data sets in order to build up a picture of the hydrogeological regime operating in the area. For example, if regions of high conductivity consistently appear up-slope of magnetic lineaments, then these lineaments (eg. dolerite dykes in Engel *et al.* (1987a)) are interpreted as acting as groundwater barriers which cause a deposition of salt on the up-slope side of the lineament.

The process of identifying and assessing all potential salt hazard sites is much more specific. A number of researchers have examined methods which can be used to predict/assess salinity risk (see for example Caccetta and Kiiveri (1996)), however, most use surface data sets such as satellite imagery which fail to examine the sub-surface causes of salinity. When geophysical data is incorporated into the prediction/assessment process, salt hazard sites are sought on the basis of specific hydrogeological models which are known to cause salinity in the survey area. For example, if salinisation is known to occur up-slope of dykes in the survey area, then first, all sites where groundwater flow intersects these groundwater barriers need to be found. Second, the hydrogeological regime up-slope of the intersection needs to be examined to determine whether the intersection poses a salt hazard, and if so, the severity of the potential hazard.

These interpretation tasks are traditionally completed manually based on visual cues. A suite of hardcopy images and maps are the interpreter's data set and tracing paper or clear film, pens, and a light table as the interpretation tools. The first stage of interpretation involves identifying boundaries and lineaments in the data set and interpreting these in geological terms. The degree of complexity in this task depends on the information being extracted from the data set. For example, under the assumption that regions of high conductivity define salt storage, the interpreter derives the salt storage map by simply tracing the boundaries of high conductivity regions off a hardcopy image. This involves a simple visual assessment of colour level for a single target variable, salt storage. By contrast, the interpretation of magnetics is much more complex and relies on a visual assessment of colour level, image texture, and extraction of lineaments, which are interpreted in terms of both lithology and structure.

The second stage of interpretation is an integration task. Interpretations from the first stage are combined to build up a more complete interpretation. The integration of data sets might confirm the existing interpretation; it might lead to new insight being added to the interpretation; or it could identify regions of inconsistency leading to a revision of the original interpretation.

Depending on the application, the final stage of interpretation will usually involve some target identification or recommendation for further action. In salinity studies, this stage involves identifying the spatial location of potential salt hazard sites based on some model (eg. intersection of saline groundwater with a barrier) and then determining the potential severity of that site.

Three key areas of inadequacy arise when the interpretation methodology just described is applied to geophysical surveys for salinity studies.

1 Whilst complex data sets such as magnetics can only be interpreted using a manual/visual approach, some of the simpler interpretation tasks (eg. deriving the salt storage map) could be more accurately and efficiently performed using a computational approach.
2 The large quantity of data available in a typical salinity study renders the integration stage of interpretation cumbersome and difficult to complete effectively. A significant risk exists that potentially important relationships between data sets will be missed using the traditional interpretation methodology.
3 The target identification process is currently missing salt hazard sites (as reported by farmers using the existing interpretation for the Broomehill district, Western Australia). A more systematic approach to identifying the salt hazard sites might solve this problem, and it would alleviate the extremely subjective nature of current salt hazard severity rating.

5. A New Methodology for Interpretation

In order to address some of the interpretation limitations discussed above, a new methodology based on a GIS platform is proposed. The GIS platform has been chosen because it offers

1. a data storage/management facility suitable for storing both raster and vector data;
2. access to a range of spatial analysis techniques which can be tailored to suit the particular requirements of this problem; and,
3. a map-making environment with which modern geoscientists will be comfortable.

The new methodology will be a four stage iterative process. It is designed to achieve a balance between the important aspects of the traditional approach (in particular, the importance of spending time familiarising oneself with the data) and the time saving and effectiveness of automating interpretation tasks in the GIS environment. The methodology, shown schematically in Figure 2, is designed to be used in applications other than salinity studies, but the discussion which follows concentrates on the salinity application. Figure 3 shows an example of the interpretation methodology applied to a typical salinity study.

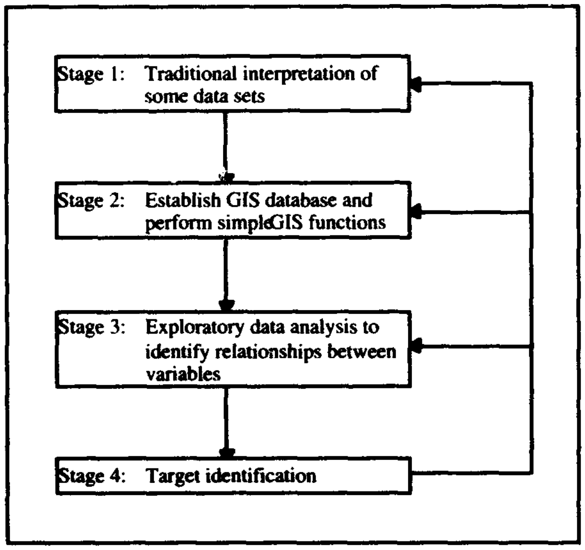


Figure 2: Schematic representation of the proposed GIS-based interpretation methodology

5.1 Stage 1

Firstly, the traditional, manual interpretation approach will still be required on some data sets, most importantly the interpretation of magnetics to produce a geology map. As noted earlier, interpretation of magnetics requires an assessment of several different image properties (colour level, texture, extraction of lineaments) which need to be interpreted in terms of both lithology and structure from the perspective of the interpreter's understanding and knowledge of the area. It is this latter component, the interpreter's knowledge, which makes automated interpretation of this data so difficult. Within the context of this new methodology, interpreting magnetics in the traditional way serves an important purpose - it allows the interpreter to become familiar with the geology of the area (Isles *et al.*, 1994). This enables the interpreter to understand the context into which results from later work can be fitted.

5.2 Stage 2

The second stage of the methodology involves establishing the entire data set in the GIS. This will involve importing data in both raster and vector formats, and may involve some digitising of data. Also, simple GIS manipulations might be used at this stage to extract simple property maps from some data sets. For example, a salt storage maps needs to be derived from the electromagnetic data. This can be obtained by slicing off the high end of the conductivity at some threshold value, dependant on the knowledge of the area. These simple GIS manipulations partially replace the tracing paper phase of the traditional approach.

5.3 Stage 3

In the third stage, relationships between multiple variables are sought. This might be done for two reasons. Firstly, in some applications, an interpreter knows that a particular combination of geophysical signatures will give him/her areas within the data set on which he/she must focus. Secondly, the interpreter will want to gain insight into how the various geophysi-

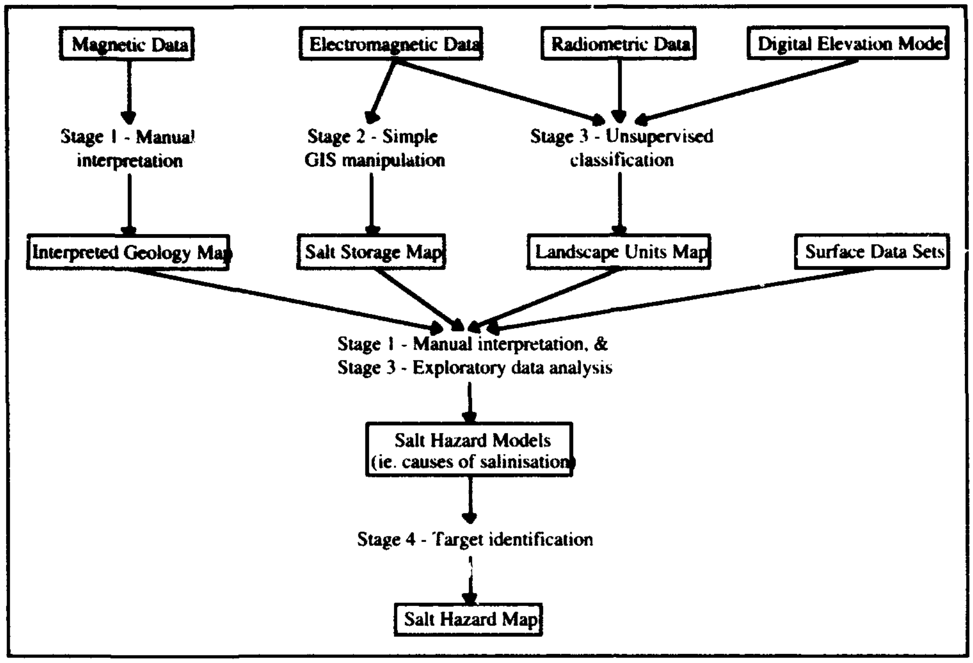


Figure 3: Schematic representation of the application of the proposed methodology to a typical salinity interpretation.

cal data sets relate to each other, and this will lead to a better understanding of the geological processes which have shaped the landscape to its present state. Exploratory data analysis, such as classification, principal components analysis, and decision tree analysis provide avenues for the structure of a multivariate data set to be elucidated. This work replaces using the light table to overlay multiple data sets, and improves on it by placing quantitative values on the relationships between variables.

5.4 Stage 4

The final stage is target identification. This is the only part of the methodology which is application specific, and its successful automation depends on the degree of complexity in the target identification process. If this methodology is to be adopted in applications beyond salinity studies, it will require a commitment of resources to translate the expert knowledge about the targets into the appropriate code. Two possible avenues exist for target identification - data driven and knowledge driven. An example of the data driven approach is decision tree analysis, where areas of known salt degradation are used to "train" the decision tree to find all other sites of salt hazard potential. This follows the exploratory data analysis of Stage 3. The knowledge driven approach requires the expert knowledge about the causes of salinisation in the study area to be translated into a series of rules. Stages 1 through 3 should have enabled the interpreter to identify the hydrogeological causes of salt degradation and define conceptual salt hazard models for the survey area. These salt hazard models can then be used to produce maps of ranked salt hazard sites for the survey area.

6. Confidence in the Proposed Methodology

In examining this proposed new methodology, the question will naturally be asked, "Why have confidence that it will work?". First, as noted previously, a GIS platform meets the technical requirements of the problem (spatial data storage, analysis, and visualisation) whilst offering a user

environment which allows the interpreter to perform digitising and overlay operations analogous to the tasks he/she performed on the light table. This should ensure that the technology transfer of this methodology is achievable.

Second, the structure of the methodology is completely analogous to the interpretation structure which the interpreter is already using. References on the art of interpretation of geophysical data are few, although a multitude report the results of interpretation. However, a set of course notes on interpretation from Isles *et al.* (1994) states the following:

As with all data sets, the interpretation should be regarded as dynamic - it will change as new evidence and ideas come to light. It is most important, therefore, to be able to retrace the interpreter's steps back to the original data so that if necessary it can be recycled. (Pg. 7)

The flexible structure of the proposed methodology ensures that this important criteria is met. In addition, the interpreter's knowledge is valued at all stages of the methodology. Researchers in the growing field of knowledge discovery consider this point to be central to the success of using computers to extract knowledge from data. Brachman and Anand (1996) state that "knowledge discovery is a *knowledge-intensive* task consisting of *complex interactions, protracted over time*, between a *human* and a (large) database, possibly supported by a *heterogeneous suite of tools*". At the centre of many knowledge discovery systems are similar analysis techniques to those suggested for this application - classification, regression, clustering, decision tree analysis (Fayyad *et al.*, 1996). The main application areas for these systems are currently large financial and health care databases which are not primarily spatial. However, the parallels between knowledge discovery and the interpretation methodology described here, suggest that the human-centred, interactive approach which has been adopted here is likely to be successful.

Finally, the spatial analysis techniques which have been selected to underpin this methodology are already widely used on similar data sets. Stage 3 of the methodology identifies classification and principal components analysis (PCA) as important spatial analysis techniques. Both of these are widely used on satellite remote sensing imagery, which is similar in many ways to airborne geophysical data. A general property of PCA is that the result demonstrates the true dimensionality of the data set, thus potentially reducing the amount of data which needs to be analysed. Also, Singh and Harrison (1985) reported that PCA applied to raw remote sensor data might yield images which are more interpretable than the original data. Both of these results would be useful in the context of interpreting airborne geophysical data for salinity. Classification, by contrast can be described as a process which transforms data into information (Jensen, 1996). In the remote sensing arena, spectral signatures for identified classes are used to tie the imagery to features on the surface (eg. different types of land cover). A completely analogous process can be used with airborne geophysical data, but spectral signatures are replaced by groups of physical properties. For example, the relationship between radiometrics, conductivity, and topography could be used to give an indication of the type of regolith. Classes with high potassium and high conductivity at the bottom of a hill would be identified as depositional areas, whereas areas of shallow bedrock would be identified by classes exhibiting very high potassium and very low conductivity on hills or slopes (pers comm G Street, 1996).

Stage 4 of the methodology refers to the use of cartographic modelling. This technique is described extensively by Bonham-Carter (1994) for use in producing maps of mineralisation potential. The approach is based on developing mineral potential models (using expert knowledge or derived from decision tree analysis) using a suite of geological, geochemical, and geophysical data. The development of salt hazard models is completely analogous to this process, although with a much stronger emphasis on geophysical data. Also, salt hazard sites will normally be specific sites, whereas mineralisation potential maps are usually regions rather than point sites. But, the success of a cartographic modelling approach to these problems, gives us confidence that this will be a successful approach to identifying salt hazard sites.

7. Conclusions

This paper has presented a new methodology for interpretation of large multivariate airborne geophysical surveys for salinity studies. However, the methodology has been constructed on principles which apply equally to interpretation of airborne geophysical data for other applications. For example, the development of a new generation of electromagnetic technology is providing mineral explorers with a new geological mapping tool. In the past, electromagnetics was used by mineral explorers to seek deep, conductive targets (likely hosts of mineralisation), but the new generation of electromagnetic systems is providing them with geological mapping information which will complement that currently obtained from magnetics and radiometrics. It is certain then that mineral explorers will soon meet identical interpretation problems to those discussed in this paper for salinity studies. This methodology provides a framework to address those problems. Resources will need to be committed to meet the development of application specific analysis modules, especially in Stage 4 of the methodology, but the overall framework is completely portable.

The strength of this methodology lies in the fact that it incorporates a "natural" approach to geoscientific interpretation with a range of spatial analysis techniques which have already proved successful in similar problems. Users of geophysical data can look forward to a future which moves beyond the use of leading edge computer-based technology for data acquisition and processing, to the use of such technology to enhance interpretation.

Acknowledgements

I would like to acknowledge the support of World Geoscience Corporation Limited who have supplied both data and funding for this project. I would also like to acknowledge the valuable assistance of my supervisors Dr Pramod Sharma, Mr Greg Street, and Dr Russell Priebennow. Finally I would like to acknowledge the Australian Key Centre for Land Information Systems (ACKLIS) who have provided the funding which has enabled me to attend this conference.

9. References

Bonham-Carter GF. (1994) Geographic information systems for geoscientists: modelling with GIS. Computer Methods in the Geosciences, Pergamon/Elsevier Science, Vol. 13.

Brachman RJ, Anand T. (1996) The process of knowledge discovery in databases. In Fayyad UM, Piatetsky-Shapiro G, Smyth P, Uthurusamy R (eds), *Advances in Knowledge Discovery and Data Mining*, AAAI Press/MIT Press, pp37-57.

Caccetta P, Kiiveri H. (1996) Assessing salinity risk on farmlands, *GIS User - The Australasian Geographic Information Systems Application Journal*, Vol. 14, pp22-24.

Duncan AC, Roberts GP, Buselli G, Pik JP, Williamson DR, Roocke PA, Thorn RG, Anderson A. (1992) SALTMAP - Airborne EM for the environment. *Exploration Geophysics*, Vol. 23, pp123-126.

Engel R, McFarlane DJ, Street G. (1987a) The influence of dolerite dykes on saline seeps in south-western Australia. *Australian Journal of Soil Research*, Vol 25, pp 125-136.

Engel R, McFarlane DJ, Street GJ. (1987b) Using geophysics to define recharge and discharge areas associated with saline seeps in south-western Australia. In Sharma ML (editor), *Groundwater Recharge*, A Balkema Publishing Company, Rotterdam, pp25-39.

Fayyad UM, Piatetsky-Shapiro G, Smyth P. (1996) From data mining to knowledge discovery: an overview. In Fayyad UM, Piatetsky-Shapiro G, Smyth P, Uthurusamy R (eds), *Advances in Knowledge Discovery and Data Mining*, AAAI Press/MIT Press, pp1-34.

Gourlay R. (1996) Mapping soil properties. *GIS User - The Australasian Geographic Information Systems Application Journal*, Vol. 15, pp30-32.

Isles D, Valenta R, Cooke A. (1994) *Interpretation and Structural Analysis of Aeromagnetic Data*, World Geoscience Corporation Ltd and Victorian Institute of

Earth and Planetary Sciences Course Notes, May 25-27, 1994.

Jensen JR. (1996) *Introductory Digital Image Processing: A Remote Sensing Perspective*, Prentice-Hall, Inc, New Jersey, USA.

McNeill JD. (1980) *Electrical conductivity of soils and rocks*, Technical Note, TN-5 Geonics Ltd, Mississauga, Ontario.

Roberts GP, Duncan AC, Anderson A, Roocke PA. (1992) SALTMAP - a new airborne electromagnetic system for salinity studies. *7th ISCO Conference Proceedings*, pp745-750.

Singh A, Harrison A. (1985) Standardised principal components, *International Journal of Remote Sensing*, Vol. 6, pp883-896.

Street GJ, Engel R. (1990) Geophysical surveys of dryland salinity, In Ward, SH (editor) *Geotechnical and Environmental Geophysics, Volume II*, Society of Exploration Geophysics, Tulsa, pp187-199.

Street GJ, Roberts GP. (1994) SALTMAP - high resolution airborne EM for electrical conductivity profiling, *Resource Technology '94*, pp544-551.

Williamson DR. (1990) Salinity - an old environmental problem. In *Year Book Australia 1990*, Australian Government Publishing Service, Canberra, pp202-211.

Wood WE. (1924) Increase of salt in soil and streams following the destruction of the native vegetation. *Journal and Proceedings of the Royal Society of Western Australia*, Vol. 10, pp35-47.

Meta Information Concepts for Environmental Information Systems

Ralf Denzer, Reiner Güttler
Hochschule für Technik und Wirtschaft des Saarlandes
Goebenstr. 40
661i7 Saarbrücken
Germany
{denzer,guettler}@htw.uni-sb.de

Presented at the second annual conference of GeoComputation '97 & SIRC '97, University of Otago, New Zealand, 26-29 August 1997

Abstract

The paper gives an overview of a number of aspects of the meta information discussion for Environmental Information Systems (EIS) over the past 7 years. While meta information has mostly been mentioned in the context of Environmental Data Catalogues (EDCs) and/or Catalogues of Environmental Data Sources (CDSs), our group uses meta information for the integration of environmental data into environmental networks. From this viewpoint, we also need EDCs and network navigation components, but our goal was one step further than the above mentioned projects: they usually stop in front of the data source and do not offer integration concepts to connect the data source into a network (Denzer, 1995).

In this paper, we will discuss a number of applications of different meta information models, which can be described by a general model to represent meta information. The generic idea of this model has been published(Denzer 1996). The first chapter is a modified extract of this publication in order to make clear the different implementation presented in later chapters.

1. A Generic Meta Information Model

In order to describe general meta data categories, we destinguish between *semantics, syntax, structure, navigation, history and summaries*. We will describe these categories with an example from the bottom up.

1.1 Semantics

By *semantical meta information* we denote additional information (additional to the raw data) which is used to describe the meaning of information. Semantical meta information is therefore the information which is needed to describe a data item such that it is interpretable by a user (from the same application field) who has not sampled the data himself.

As a less abstract term we can also use the term *data description* as a synonym for *semantical meta information*.

As you can see in Fig. 1, we append a set of meta information items to the raw data. The meaning of the meta information items can be general knowledge (like *address of data provider*) and therefore be understood by the general public, or it can have domain specific meaning which is only understood by an expert in the specific application area (like *field method*). This means that the set of meta information items may be different for different user groups.

1.2 Syntax

By *syntactical meta information*, we denote information which is used to describe the way the raw data is stored and/or can be accessed. Syntactical meta information is unimportant for end users and is only used by software systems to access the data. Syntactical meta information usually consists of information about the data type of the raw data and an access method.

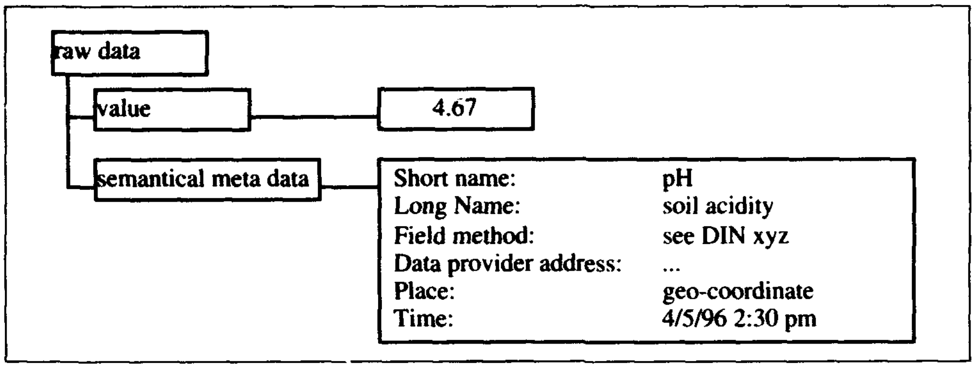


Fig 1. Semantical meta information

Fig. 2 shows how syntax information is appended to the existing set of information describing the raw data. A minimum of information regarding the data type and access methods may be given depending on the way the data is stored (in this example a relational database).

1.3 Structure

Up to this point, we have shown single data items and how their semantics and syntax can be described. In reality, data objects can_nott be described as single items very often. Commonly, aggregates of data items form an *environmental object*, and there is meta data which applies to the whole object as well as to the single raw data item. Therefore it is necessary to describe the *structure* of data objects as well, we denote this description as *structural meta information*.

In fig. 3, several data items with their meta data are composed to an object. Additionally, a semantical description of the overall object is given, which consists of the meta information applying to each of the items (e.g. *data provider address* would no longer be meta information of *pH*, it would be part of the semantical description of the object).

The semantical description of the overall object is again a list of meta data items, according to the description of a single data item. In this case, the description of structure is such that an *object consists of a list of attributes. It is important* to notice that an attribute can be of type *datatype* (single datatype, vector, time value, ...) or of type *object* itself. This also applies to each of the metadata items (an object of class *field method* is meta data item for an at-

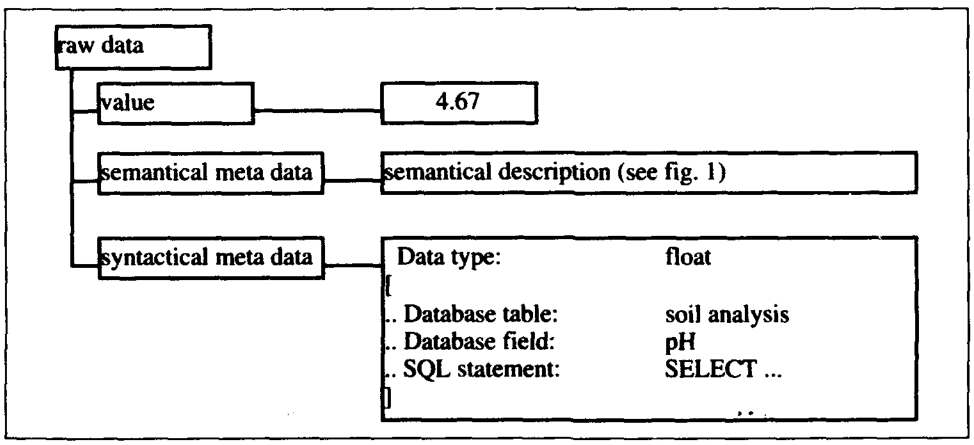


Fig 2. Syntactical meta_information

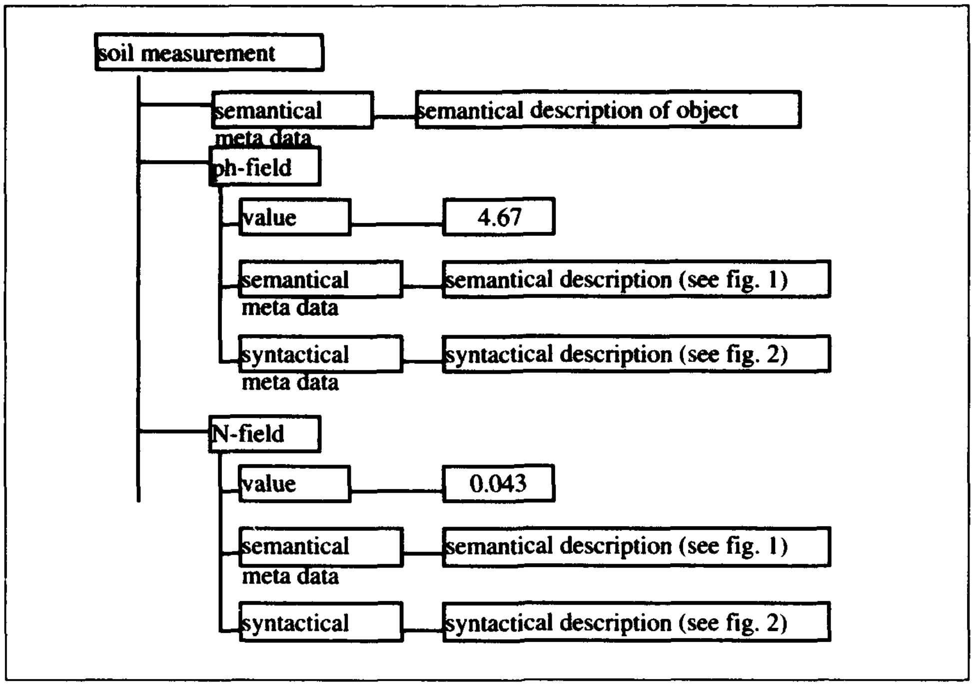


Fig. 3. Structural meta information

tribute *pH* of an object from class *soil measurement*). This meta data model is therefore inherently object oriented, but it would also be possible to describe the structure of the whole with other methods.

1.4 Navigation

Semantics, syntax and structure are entities used to *describe* environmental objects. Another important issue is to *locate* environmental objects. By *navigational metainformation* we denote such information which is used to locate objects and data sources of interest. Navigation occurs within systems (search masks, keyword lists, inventories, etc.), or among systems or even whole networks. Environmental *data catalogues* are one of the means to locate objects.

Fig. 4 gives an example of a data catalogue for one information system. The catalogue combines a list of object classes, a hierarchical tree (table of contents) and links from chapters to class descriptions. Such a catalogue may also include a list of keywords which can be inspected to find information of interest. A data catalogue on the level of an organization or of a whole network would look differently (see Fig. 5). The entries in the table of contents or keyword lists build links to information systems (e.g. data sources). We call such a catalogue *meta catalogue*.

Navigation is much more than that. It includes issues of search engines, statistical information and it raises issues about how to organize information sources over a whole network. Bad experiences with information searches on the Web illustrate these problems.

1.5 History

The problem of history of environmental measurements has widely been ignored over the past years. Why is this the case? First, history means that samples may be produced *by different measurement technology* over the course of time. This increases the design and maintenance efforts for an information system significantly . Second, history also means that the *data structures* change over time. This is even worse for an information system design. Third, his-

tory, or changing methods, produces problems in the comparability of data, which is a problem for the scientists, a problem they produced themselves by changing the method.

History in terms of meta data means that historical records about how data has been sampled are important for the description of the data. In practice this means that each of the meta data objects in Figures 1 to 3 must be recorded historically and may even change their structure.

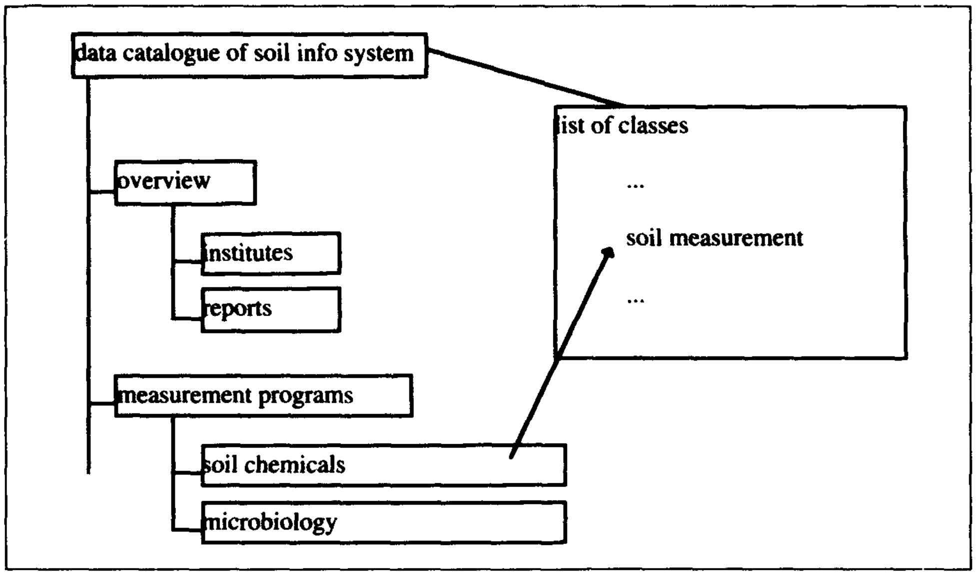


Fig. 4. Environmental data catalogue for an information system

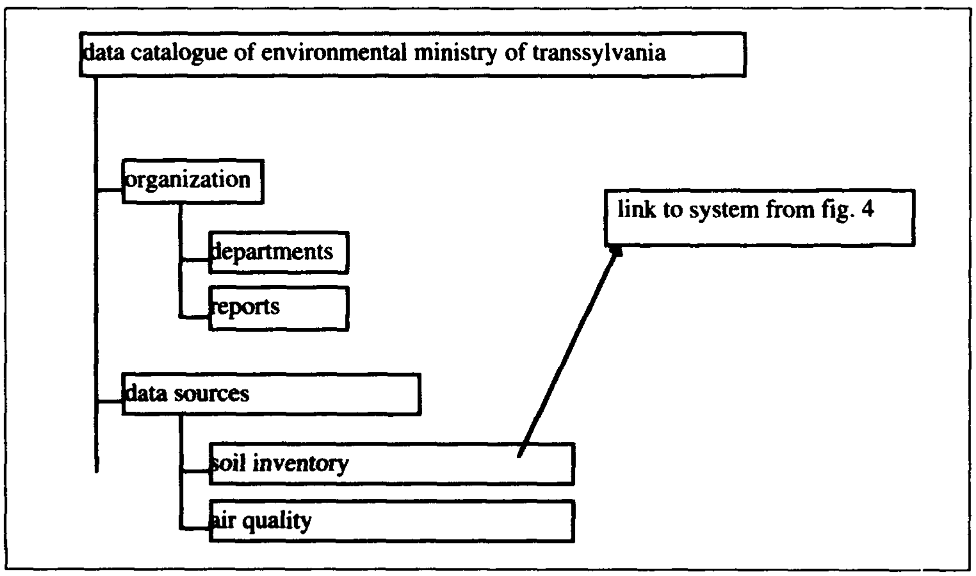


Fig. 5. Environmental data catalogue (meta catalogue) of an organization or network

1.6 Summaries

Summaries are used to give an overview at each level of meta information (e.g. number of classes in a catalogue, number of objects in a class, percentage of attributes used, overall summary of time or geographical scale, etc). Summaries are implemented to help users during navigation and although these are very simple mechanisms, they are not used frequently.

2. *Application Examples*

In this chapter, we give three examples of the implementation of the concepts mentioned above. The examples are very different in nature.

2.1 SIRIUS Meta Information Model

The SIRIUS (Saarbrücken Information Retrieval and Interchange Utility Set) system was our first implementation of a meta information concept. The goal of SIRIUS is to provide an integration architecture for open EIS. This architecture has been documented in various publications (Denzer, 1995). Meta information in the context of SIRIUS is mainly used for the following purposes:

- to provide a data catalogue for an existing information system,
- to document the information classes of this information system in terms of class syntax, structure and semantics,
- to use the class documentation for the access of the information system and, to provide networked catalogues (meta catalogues) for the organization of a SIRIUS network.

The meta information used in SIRIUS is a very simple model, where

- catalogues are hierarchical trees,
- classes are described by a set of attributes,
- classes are linked into nodes of the catalogue,
- the class structure is given by a list of primitive attributes,
- the attribute syntax is given by its data type (plus an optional list attribute),
- each class can have a different description, and
- the semantical meta information for each class and each attribute is just a free text.

In terms of distribution of data, SIRIUS is completely network transparent.

2.2 FAM Meta Information Model

FAM (Forschungsverbund Agrarökosysteme München) is a big agricultural research project funded by the German government. In this project the operation of a farm is monitored on a long term time scale. A large number of institutes (at the time of our involvement in the project around 60) collect all possible data associated with the operation of this farm and use this information for ecological assessments.

In 1994 and 1995, our group developed a meta information model for the database of the FAM project. This model is, to the best of our knowledge, the most detailed and flexible meta information model implemented at this time. The differences between the FAM model and SIRIUS are twofold:

- no network component (which was not needed), and
- the description of classes is much more detailed.

Compared to SIRIUS, the FAM model describes classes as follows:

- a class has again a list of attributes, but these attributes can be of any type, including other objects, therefore the data model is recursive
- each class and each attribute has a set of meta information attached to it, and this set is not only a free text but a list of meta information attributes which can be of any type (primitive types and objects); as objects can contain objects, also the meta information model attached to every class and/or attribute can be recursive
- the meta information contents of any class or attribute can change over time, reflecting change in the reality (i.e. the model can store a history of e.g. measurement instruments)
- the class structure and the meta information struc-

ture (of a class or any attribute) can change over time (i.e. even if the reality changes its structure, the information model can reflect this and can even remember the structure at a certain date in the past).

It is easy to imagine, that this meta information model is neither trivial to understand nor trivial to implement. But our investigations showed clearly, that this is the set of information needed to document a long scale research program such that the information can still be used after a longer time period.

2.3 TEMSIS Meta Information Model

TEMSIS (Transnational Environmental Management Support Information System) is a project funded by the EU under the Environmental Telematics program (Schimak 1996). The goal of the system is the support of environmental information and planning in the area around the French-German border near Saarbrücken and Saargemuines. We are part of a consortium of 8 partners developing this system.

Our colleagues at the Austrian Research Center Seibersdorf are developing the meta information server. Our group is responsible for the information services between the server and data sources as well as between server and client applications on both sides of the border (for this purpose, a port of SIRIUS is used). As our tasks in the overall project were related, we have worked closely together in the modeling of the meta information. The TEMSIS meta information model is located between the two models mentioned above in the following areas:

- meta information is a list of primitive data type objects, not a text as in SIRIUS, but not recursive as in FAM
- the TEMSIS model does not distinguish between objects, attributes and classes. What this really means compared to SIRIUS or FAM will come out in the future. It seems depending on the way the meta information is organized in the catalogue, this can be completely irrelevant to the end user and will only be noticed by the system designer.
- the TEMSIS model does not store any history, nor does it directly reflect the distributed nature of EIS (TEMSIS uses one centralized server)
- the TEMSIS model introduces a new very powerful idea from our friends in Seibersdorf. Links between information objects, which are used to model relationships between objects and can be used extensively for navigation.

3. Discussion

The three models are very different in nature and purpose. They also reflect the different reasons of how and why to use meta information in an EIS. SIRIUS uses a very simple model for the interconnection of EIS and therefore describes objects only on a very abstract level. FAM, in comparison, is an extremely detailed and sophisticated model, which is able to model anything, but it is not easy to use. We have not been able, due to limited funds in this particular project, to implement the user interface components which handle the complexity of the model, especially for the persons who have to maintain the meta information system. The TEMSIS model appears to be a good compromise for a public information system, which does not have the same detailed need for documentation as is found in a research program. However, we do not have any experiences yet with the model, as the demonstrator system will be installed this summer. Also, it is limited to one central server, although the information services are capable for link up to a network.

4. Conclusion

The comparison of the three projects shows that there is not THE meta information model for the world or for EIS. In every case and under different circumstances, a different way to use meta information will be useful. But we strongly believe that there is a generic way of thinking about meta information, which may be reflected by the first chapter of this article and which may have been implemented in a most generic way in the FAM meta information model.

If we look back to the past 7 years, since the strange word "meta information" became common (and not many peo-

ple know what it is), we can also see that a convergence towards usable approaches in EIS took place.

References

Denzer (1995) R. Denzer, R. Güttler, SIRIUS - Saarbrücken Information Retrieval and Interchange Utility Set, in: R. Denzer, D. Russell, G. Schimak, International Symposium on Environmental Software Systems 1995 (ISESS 1995), Malvern, PA, USA, June 1995, Chapman & Hall, 1995

Denzer (1996) R. Denzer, R. Güttler , Requirements of Meta Information Models for the Environmental Domain, Journal of Computing and Information, Vol. 2, No. 1, pp. 1112-1120, 1996

Schimak (1996) G. Schimak et. al. (TEMSIS consortium), TEMSIS - a first step towards transnational environmental information systems, ECO-Informa 96, Global Networks for Environmental Information, November 1996, Epcot Science and Technology, Lake Buena Vista, Florida, USA

Computation and visualisation of historical geographical data for acoustic channel modelling

R.M.Dunbar, L.M.Linnett, Y.Rzhanov, J.Bell
Heriot-Watt University,
Edinburgh, UK
rmd@cee.hw.ac.uk

Presented at the second annual conference of GeoComputation '97 & SIRC '97, University of Otago, New Zealand, 26-29 August 1997

1 Abstract

The paper describes how historical geographical data in the form of depth soundings have been used to improve the understanding of hydroacoustic signal propagation through a large ducted water channel. The geo-computational elements of the paper fit within the larger framework of research into underwater vehicles, subsea communications and imaging, carried out by the Ocean Systems Laboratory over the past 25 years. The paper includes a brief review of these activities and of the original hydrographic survey of Loch Ness, carried out around 100 years ago. The method of digitising the original data and the production of 3D static and moving visualisations is then discussed in the context of acoustic channel modelling, and the paper concludes with an outline of continuing work in relation to simulated test environments.

2 Introduction

2.1 Theme of Paper

The purpose of this paper is to describe how historical geographical data has been used to give a clearer understanding of actual observed effects in relation to the modelling and experimental validation of underwater acoustic signals. Depth soundings which were collected meticulously 100 years ago are believed to be a reliable data set for this study and visualisations which have been produced recently have in fact explained certain anomalous signals. The application of these basic geo-computational techniques is proving to be of considerable value in the generation of simulated test environments for on-going research.

2.2 Background

Research activities in Underwater Technology at Heriot-Watt University began in 1969 following a survey carried out to identify a totally new research direction for the Department of Electrical and Electronic Engineering (Dunbar, 1970). Research studies were initiated into subsea vehicles, instrumentation, viewing, communication and navigation, and activities were focused on a major project which had the objective of designing, building and operating Scotland's first remotely operated vehicle (ROV) system. The first ANGUS vehicle was successfully tested in deep water in 1973 (Dunbar, Holmes, 1975) and the ANGUS 002 and 003 vehicles followed in 1976 and 1979, in the development of ROV systems with automatic control and navigation (Russell, Dunbar, 1990).

From 1976 studies expanded into tetherless vehicle systems (now 'AUVs', Autonomous Underwater Vehicles) and this forced the development of through-water communication systems (Dunbar, Carmichael 1990), sonar systems, and sub-surface video transmission and bandwidth compression techniques (Dunbar, Settery, 1985). ROV and AUV trials were carried out in test tanks, harbour areas, from ships at sea and in Scottish lochs, and it was during experiments in Loch Ness that a World War II Wellington bomber was located on the bottom of Loch Ness and eventually

recovered in 1985 (Holmes, 1991). More recently, a European Community Marine Science and Technology (EC-MAST-II) project managed by the University used Loch Ness as one of the test locations.

The project (1992-1996) was entitled European Experimentally Validated Models of Acoustic Channels ('EEVMAC') and it had as its prime objective the precise measurement and data logging in absolute terms of hydroacoustic signals in the 2kHz to 80kHz range, with various modulation formats, transmitted over various ranges underwater, together with associated oceanographic and environmental parameters. The data would then become available for the validation of acoustic channel propagation models (Dunbar, McHugh et al, 1994).

3 Hydroacoustic communications and modelling

Modelling the path, the spreading loss, and the attenuation of hydroacoustic signals is a complex process, particularly for regions with multiple boundaries which lead to multipath propagation, and many models have been developed with various degrees of precision (Buckingham, 1992). Model development is often application driven: for instance, modelling of the multipath environment in a search for automatic methods of cancelling multiple echoes in a time-varying environment (Dunbar, Carmichael, 1989); and the development of models for the simulation of synthetic sonar images (Bell, 1995), to aid the interpretation and classification of sonar and sub-bottom seismic images (Linnett, 1991).

Mathematical models and simulations require real test data for their validation and correction, and the gathering of such data under carefully controlled conditions has been carried out successfully within the EC-MAST-II project 'EEVMAC', mentioned above. Currently, work of a similar nature is being carried out within the EC-MAST-III project 'PROSIM', which is an impulse signal variant of EEVMAC.

4 Original hydrographic survey of Loch Ness

4.1 Historical document

A bathymetric survey of Scottish fresh-water lochs was carried out by Sir John Murray and Laurence Pullar over the years 1897 to 1909 (Murray, Pullar, 1910), and consequently the year of this present conference has particular significance. The survey was an outstanding scientific achievement and a reading of the original documents leaves one with a sense of admiration and respect for the investigators when one considers the scale and precision of their measurements in the light of the experimental equipment at their disposal. To quote from the introduction to their report:

"During the course of the Lake Survey work 562 of the Scottish fresh-water lochs were surveyed.....all lochs were surveyed on which boats could be found at the time the work was being carried out.....To transport a boat to many of the remote lochs in the Highlands would have entailed much labour and difficulty, not to speak of the objections of proprietors, keepers, and others, who do not wish to have grouse moors and deer forests disturbed at a time of year when the lochs are most accessible."

It was an immense undertaking, which included in addition to the depth soundings, observations and measurements relating to topographical, geological, physical, chemical and biological features.

4.2 Method of survey

For deepwater lochs the *'F.P.Pullar sounding-machine'* was employed. This was a well designed and engineered mechanism which included a drum containing over 1000 feet of three strand galvanised steel wire which passed over a pulley having a circumference, to the centre of the wire, of precisely one foot, and a group of measuring dials recording feet, tens of feet, and hundreds of feet as the sounding weight was lowered to the bottom. It was thus possible to make precise depth measurements without difficulty. It was more complicated to determine the position of the soundings. Various methods were tried but *"it was found that the*

most accurate method was to take the soundings as quickly as possible while rowing across the loch from one point to another. The soundings were taken, say, every thirty strokes of the oars, and the total number of the soundings was placed equally along the line, thus distributing any errors". The method was found to be "*extremely accurate for long, narrow lochs*", of which Loch Ness is a prime example.

In the case of Loch Ness, over 1000 soundings were taken during the course of 79 across-loch transects. On completion of the survey the soundings were transferred to 6-inch Ordnance Survey maps of the area. Later, clean tracings were plotted on cloth and contour lines of depth were drawn in at equal intervals. These original tracings later became the source data for an Admiralty chart of Loch Ness which continues to be published as chart number 1791. Additional soundings of a small area at the North end of the loch were taken in 1918, to provide greater detail near the entrance to the Caledonian Canal; however, the 79 transects remain the primary data archive.

4.3 Current survey technology

More recently, various sonar surveys of the loch have been carried out, the most comprehensive and detailed to date being undertaken during the course of 'Project Urquhart'. In July 1992 a survey vessel employing a state-of-the-art multibeam swathe echosounder carried out a detailed sonar survey of the loch and as a result 3-dimensional views of the loch were computed and publicised (Witchell, 1992). Discussions are underway with the organisation holding the sonar and sub-bottom seismic data with a view to comparing and developing the two approaches to 3-dimensional visualisation.

5 *Examination and formatting of the original data*

Copies were obtained of the original maps of Loch Ness, bearing the actual soundings across the 79 transects, and by enlargement were overlaid on a montage of current 1:25000 scale Ordnance Survey (O.S.) maps of the area. By comparison between the old and new maps, the modern O.S. co-ordinates were deduced for the shore (zero level) ends of each transect. Then, by linear interpolation, the equivalent O.S. easting and northing grid points were computed. The interpolation was based on the assumption that the soundings were equally spaced, the same assumption as made by the original surveyors. An example of such interpolated data is given below, for the first transect from the SW end of Loch Ness, near to Fort Augustus.

Sounding	Easting	Northing	Depth (ft)	Depth (m)
Nth shore	2 384 000	8 092 000	0	0
1	2 384 400	8 091 400	3	0.91
2	2 384 900	8 090 900	94	28.65
3	2 385 300	8 090 300	160	48.77
4	2 385 800	8 089 800	227	69.19
5	2 386 200	8 089 200	250	76.20
6	2 386 700	8 088 700	241	73.46
7	2 387 100	8 088 100	207	63.09
8	2 387 600	8 087 600	56	17.07
Sth shore	2 388 000	8 087 000	0	0

All 79 transects were examined in this way and 79 [x,y,-z] data files based on absolute O.S. co-ordinates were produced for use in subsequent analysis.

6 *Generation of 2D and 3D images*

6.1 Preparation of data

To improve interpolation between lines of soundings the 79 data files were padded with zeroes at points beyond the N and S ends of the transects, to aid the performance of a 'N nearest neighbours' algorithm. The positions of the augmented transects are shown in figure 1. Although it would have been possible to create [x,y,z] files to include *above* waterline elevations by inspection of the original maps, the intention was to merge the historic sub-surface data files with contemporary O.S. digital data files of surrounding terrain.

6.2 Computation and visualisation of data

The original height data refers to a narrow region which traverses from South-West to North-East and therefore within the bounding rectangle of Loch Ness the data points in total are very sparse. To simplify the computations the co-ordinate system was rotated by 38 degrees anti-clockwise, to produce in effect a vertical bounding rectangle. Prior to rotation each data point consisted of an easting

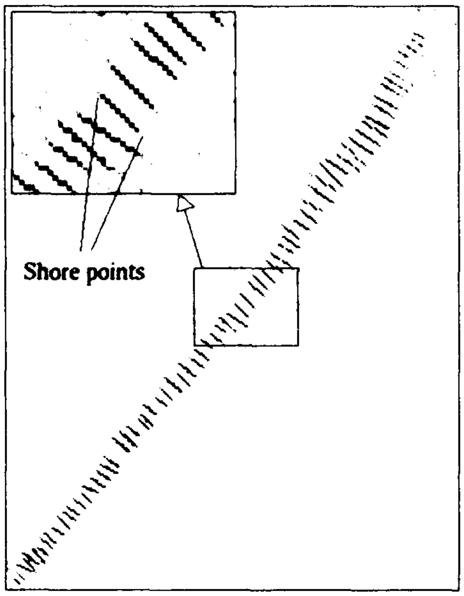


Figure 1: Original transects with augmented zeroes

and northing and depth value. The data points were not uniformly distributed so they were interpolated to achieve a uniform distribution which produced a three-dimensional scene consisting of 100m x 100m grid squares. This produced a coarse rendition so the image was scaled by a factor of 5 to produce 2.8 million polygons. This then produced an image of size 675 x 2100 pixels representing an area of 13 kilometres by 42 kilometres. The contemporary O.S. digital survey data is gridded at 50m x 50m intervals and spot heights in metres at these intervals are given. The two data sets were merged after alignment, the O.S. survey data thus providing the landscape surrounding Loch Ness. A straight line two-dimensional rendition of the two data sets is shown in figure 2, where the water level contour is the common factor between the two plots.

6.3 Interpolation of data

To interpolate values in the grid points the standard method of N nearest neighbours was used (Davis, 1988). This approach was used primarily because of its relative efficiency on dealing with the irregular and sparse nature of the original data set. For each grid point (k), N nearest points from the data set are found (N is usually in the range 3 to 6), and the height in the grid point is calculated as a weighted average:

7 *Production of a 'fly- through' of Loch Ness*

The 3D scene was rendered using a custom- designed program on a Silicon Graphics 02 computer. Commercial packages were considered but data handling software under development by the research team for other image interpretation applications proved to be more cost effective

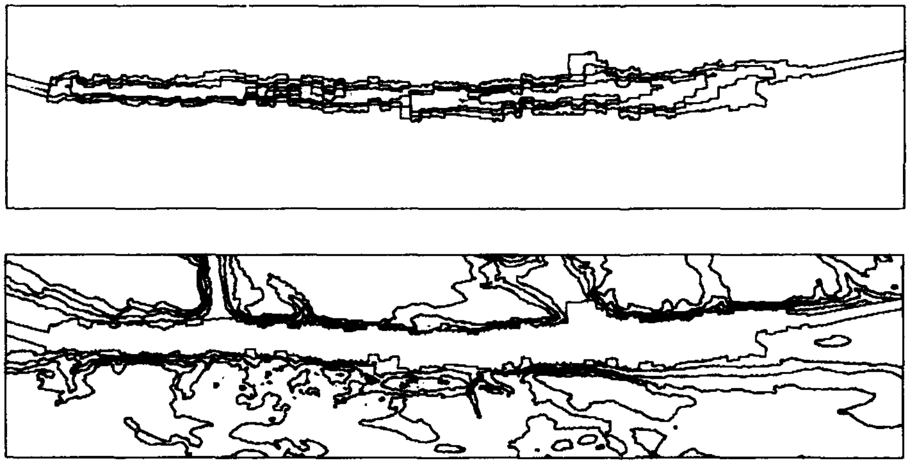


Figure 2: 2D rendition of Loch Ness: sub-surface contours (upper), surrounding terrain (lower)

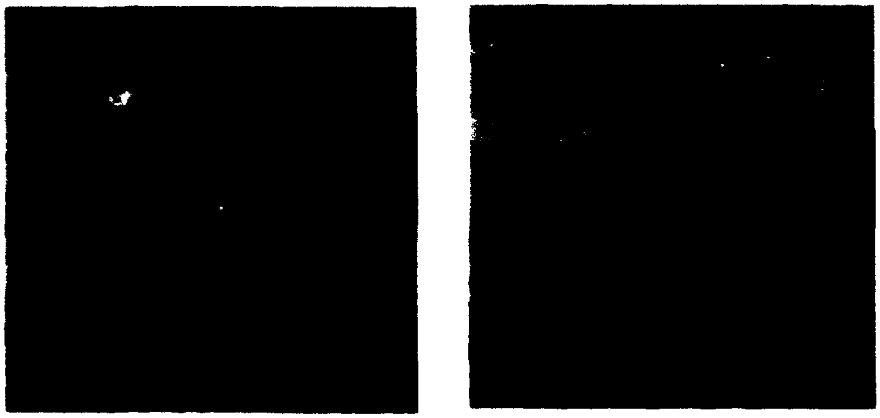


Figure 3 3D renditions of Loch Ness using combined data sets, for southern and central sections of the loch

and flexible in manipulating data sets from different sources. The system requires an "observer" position and a "look-at" position. From this data the perspective view is computed and the image is saved as a single frame. This process is repeated about 2000 times to produce a movie. For each frame the "observer" and "look-at" positions are changed slightly to create the impression of movement. These effects may be observed in a video sequence, which will be presented at the conference.

8 *Application to acoustic channel modelling*

8.1 Ray tracing model

In trying to predict the characteristics of an underwater acoustic signal that might be received at a particular depth in the water column, and at a particular range from the transmitter, a ray-tracing model is helpful. The path that a particular ray will follow is a function of the sound velocity in regions of water through which the ray passes, and in turn, the sound velocity is a function of temperature, salinity (or electrical conductivity) and pressure. Since these last three parameters are not constant, one must have knowledge of their spatial (and temporal) distribution in the three dimensional region of interest in order to make a sensible estimate of the ray path. Consequently, a common measurement made in underwater acoustics is a 'CTD' (Conductivity, Temperature, Depth) profile, and the more CTD profiles that are available along the route of the acoustic pressure wave, the greater is the precision with which one can predict its path. As a result of the variation of sound velocity with depth, sound ray paths are in general curved, and a typical ray plot will illustrate computed ray paths for a selection of launch angles. Examples of such ray paths are shown in figure 4 where ray plots have been computed for Loch Ness using typical measured CTD values. Ray tracing models become sophisticated when absorption, reflection and scattering coefficients at surface and seabed, seabed stratification and topography, and frequency dependent attenuation are also taken into account (Bell, 1995).

8.2 Visualisation of 3D acoustic environment in Loch Ness

Ray tracing models as illustrated above can predict multipath signals and over path lengths of several kilometres it is normally observed that multiple reflected signals arrive very shortly after the direct path signal, if the distance/depth ratio is large, which was the case for Loch Ness. However, unlike similar hydroacoustic trials carried out in the Mediterranean Sea, in an open area with a fairly flat seabed, in the case of Loch Ness multiple signals were observed after an unexpectedly long delay.

During field trials in 1995 a set of 1.9kHz ASK test transmissions were made over a path length of 7 km, the acous-

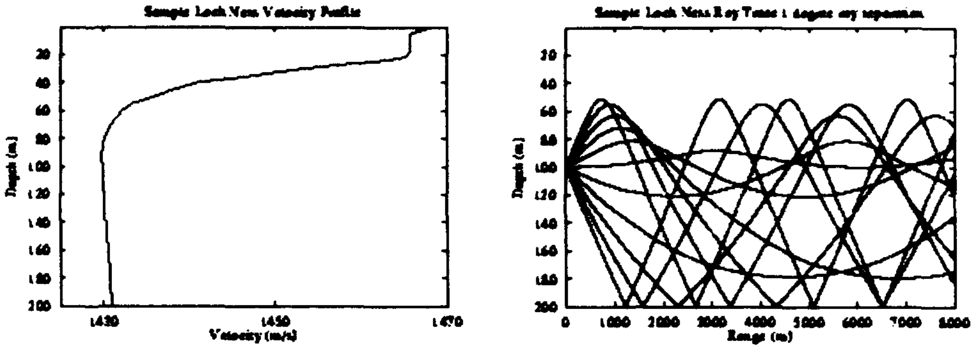


Figure 4: Typical sound velocity profile and ray trace for Loch Ness

tic projector and receiving hydrophones being at mid-water depth. These tests were performed along the centre line of the SW section of Loch Ness, in a stretch of water where the loch width was approximately 1.5 km. For a water depth of 200m and a direct path length of 7km, first and second multiple signals could be expected to arrive within 20ms of the direct ray, and this was observed to be the case. A additional group of signals was observed however with a delay of approximately 100ms. A short calculation as follows indicates that these signals are likely to be due to reflections from the sides of the loch, as visualised notionally in figure 5.

The path length for a mid-water signal undergoing a single reflection in the horizontal plane in such a location would be 7.16 km, as compared with a straight line path of 7 km. Consequently, the difference in propagation times would be the path difference divided by the speed of sound, i.e. 160m /1430 m/s = 0.112 s. When recordings of the test signals were examined in detail in the region of 0.1s from the start of the received pulse, evidence was found of significant constructive and destructive interference on the signal, a characteristic of multipath signals. An example of such a signal is illustrated in figure 6. There was also the normal evidence of surface and bottom reflections, earlier in the received signal, but reflections with such a delay and magnitude would have been unusual for open water situations.

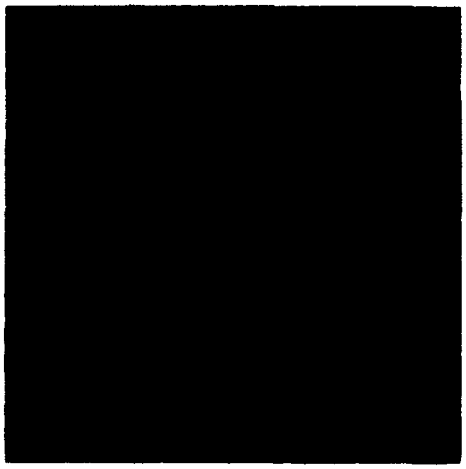


Figure 5: Notional visualisation of direct and side-reflected ray paths

9 Conclusions

It is believed that geographical visualisation adds a dimension to acoustic modelling that considerably enhances the understanding of the overall communication or sonar process. Consequently, the technique is being further developed for more detailed analysis of existing acoustic data, from Loch Ness and other test sites. Moreover, the experience gained through this present investigation suggests that the 3D visualisation is a valuable framework for a simulated test environment, where signals from multiple sensors may be fused and observed.

By setting up a flexible geographical framework in which to insert data as it becomes available the investigator has a

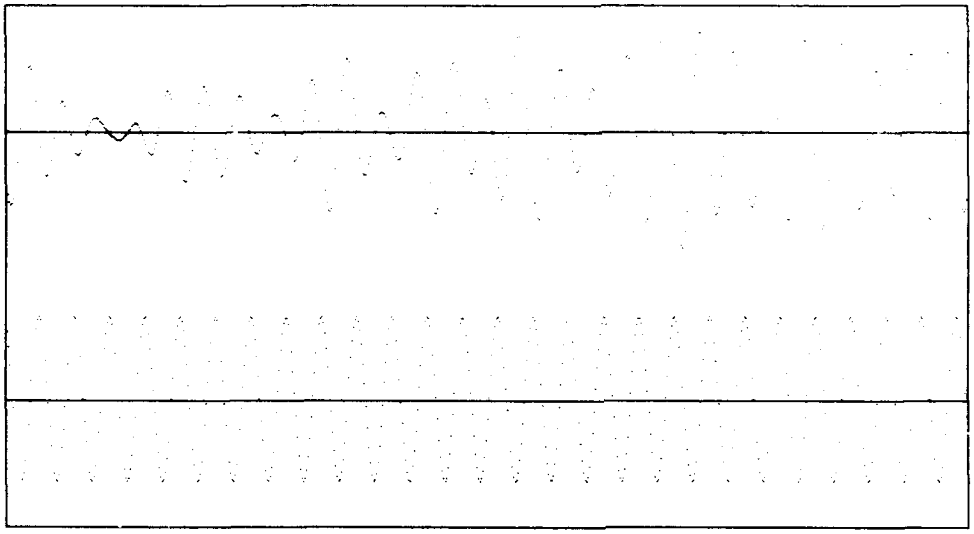


Figure 6: First arrival of signal reflected from side(s) of Loch Ness. lower trace is reference

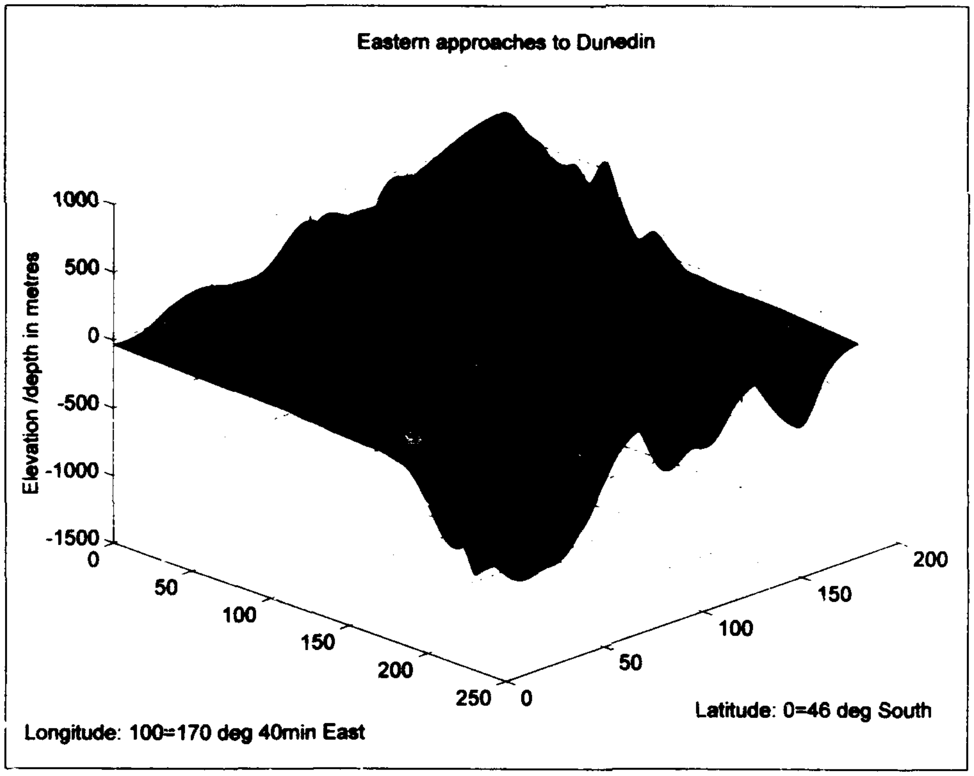


Figure 7: Visualisation of approaches to Dunedin, from limited data set.

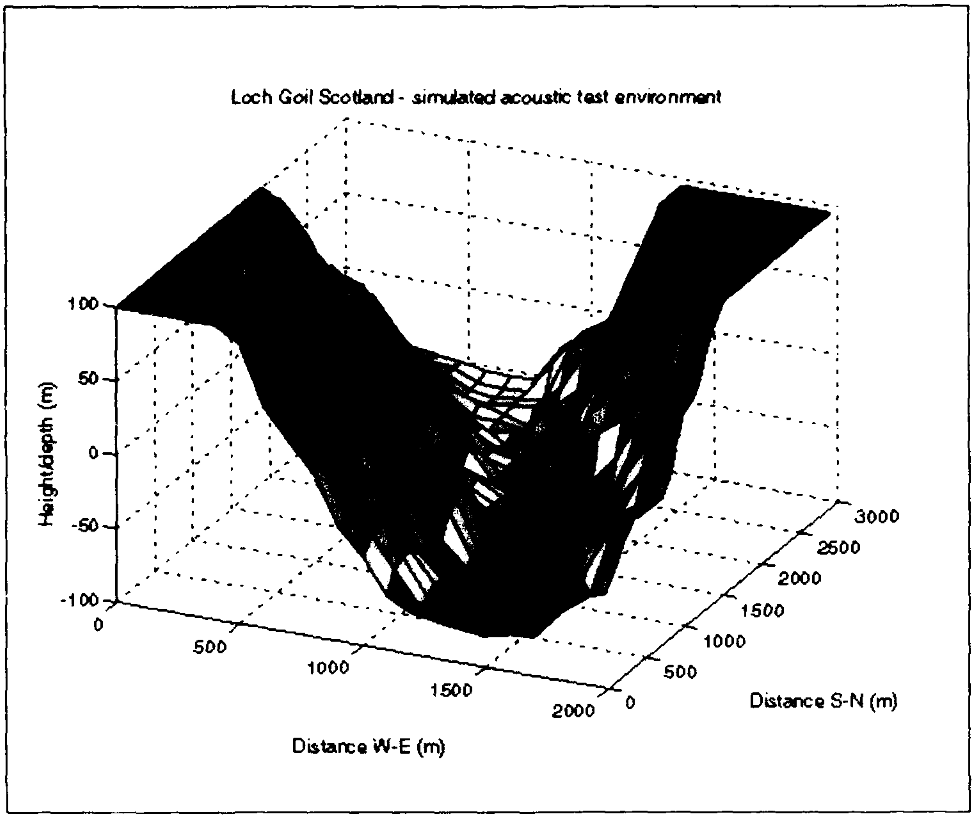


Figure 8: Visualisation of Loch Goil acoustic channel.

powerful mechanism for data logging and analysis (Dunbar et al, 1990). As an illustration, figure 7, the Eastern approaches to Dunedin have been examined on an Admiralty chart and a 3D perspective view produced under MATLAB(R), using 23 x 14 points and cubic spline interpolation, to provide a visualisation which could be used, for example, as a first step in modelling the arrival paths of ocean acoustic signals. As a further illustration, figure 8, a section of Loch Goil in Scotland has been modelled using a similar approach, and this model is currently being used to interpret hydroacoustic signals received during recent trials.

10 Acknowledgements

The authors wish to thank the National Library of Scotland and Ordnance Survey for permission to make use of historical records and digital data sets respectively, in the production of visualisations referred to in this paper. The authors also wish to acknowledge the support of the EC MAST programme, the UK Engineering and Physical Sciences Research Council, and the UK Ministry of Defence, in projects which have included field trials referred to in this paper; and the assistance of Miss Linda Purves of Heriot-Watt University in the transcription of historical data.

11 References

Bell J.M. (1995) A Model for the Simulation of Sidescan Sonar. PhD thesis, Heriot-Watt University, September 1995.

Buckingham M.J. (1992) Ocean-acoustic Propagation Models. J.Acoustique, 1992, 223-287.

Clay C.S., Medwin H. (1977) Acoustical Oceanography, Wiley-Interscience, John Wiley and Sons, New York, London, 1977, 83-106.

Davis J.C. (1988) Statistics and Data Analysis in Geology, John Wiley, 1988.

Dunbar R.M.(1970) A Survey of Underwater Technology. Heriot-Watt University, Dept. Computing & Electrical Eng., RM 70/4.

Dunbar R.M., Holmes R.T.(1975) ANGUS. J.I.E.E., Vol.21, No.2, April 1975, 433-436.

Dunbar R.M., Settery, A (1985) Video Communications for the Untethered Submersible ROVER. 4th Int Symp Untethered Sub Tech, Univ New Hampshire, 'SA, 1985, 140-149.

Dunbar R.M., Carmichael D.R. (1989) Adaptive Estimation of an Underwater Acoustic Channel using Real Test Data. 6th Int Symp Untethered Sub Tech, Univ New Hampshire, USA, 1989, 42-49.

Dunbar R.M., Carmichael D.R. (1990) Subsea Communications for Semi-Autonomous ROVs. Ocean Resources, Vol.2, Kluwer Academic Publishers, 1990, 185-189.

Dunbar R.M., Hennings I., Kullander L., Linnett L.M., Oskarsson O., Torlegord K., Shippey G., Vikgren K.J., Werner F. (1990) Monitoring of Seafloor Changes with Multisensors: Problems related to Positioning, Image Rectification and Classification. Int Conf on Submarine Systems, U-90, FOA, Stockholm, Sweden, May 1990, Vol.2, Paper 17.

Dunbar R.M., McHugh R., de Malet Roquefort S., Bathgate L. (1994) High-level Control of Automated Oceanographic and Acoustic Equipment and Experiments. OCEAN-OSATES 94, IEEE, Brest, France, Sept 1994, Vol.1, 678-682.

Holmes R.T. (1991) One of Our Aircraft: the Story of 'R for Robert' the Loch Ness Wellington. Quiller Press Ltd, London, 1991.

Linnett L.M., Clarke S.J., Graham C., Langhorne D.N. (1991) Remote Sensing of the Seabed using fractal techniques. IEE Electr and Comm J., Oct 1991, 195-203.

Murray J., Pullar L. (1910) Bathymetrical Survey of the Scottish Fresh-Water Lochs. The Challenger Office, Edinburgh, 1910.

Russell G.T., Dunbar R.M. (1990) Intelligent Control and Communication Systems for Autonomous Underwater Vehicles. International Advanced Robotics Programme (IARP), MBARI, Naval Postgraduate School, Monterey, CA, USA, Oct 1990, 1-14.

Witchell, N (1992) Sounding the Secrets of the Depths of Loch Ness. World, BBC, London, Oct 1992, 62-71.

Data Visualisation for New Zealand Forestry

Alan J. Thorn
Scientist,
Resource Monitoring Unit
New Zealand Forest
Research Institute
Private Bag 3020, Rotorua,
New Zealand

Terry C. Daniel
Professor Psychology and Natural
Renewable Resources,
University of Arizona
Environmental Perception Laboratory,
Department of Psychology
Tucson, Arizona 85721,
USA

Brian Orland
Professor Landscape Architecture,
University of Illinois
Imaging Systems Laboratory,
214 Mumford Hall,
1301 West Gregory Drive,
Urbana-Champaign,
Illinois 61801, USA

Presented at the second annual conference of GeoComputation '97 & SIRC '97, University of Otago, New Zealand, 26-29 August 1997

Abstract

The New Zealand Resource Management Act emphasises, among other things, the evaluation of the effects of forestry operations, and the inclusion of those affected into the decision making process. Visual images are a useful method for displaying the effects of planned forestry activities, and are easily understood by most of the general public. While the creation of fully accurate photo-realistic images is still the domain of super computers, it is possible to come close using data visualisation techniques that have been developed for a desktop computer.

The data visualisation techniques reported in this paper focus on the creation of photo-realistic, oblique view images depicting the predicted results of alternative management activities. A Geographic Information System (GIS) is used to develop a digital terrain model of the scene. Other information from the GIS database, such as forest stand boundaries, is shown on or draped over terrain model. Biophysical models are used to 'grow' the trees to be placed in the landscape using software called SmartForest II. Calibrated analytical images are positioned on the terrain model to match the planned forestry activities. This creates representations that are sufficiently accurate in all dimensions, and facilitates rendering photo-realistic images. The resulting images have been used in surveys to gauge public preference of forestry options.

Background

A successful forest industry based on intensively managed *Pinus radiata* plantations has been part of the New Zealand agricultural economy for over 30 years. Currently the 1.5 million hectares of plantation forest is the third major contributor to the New Zealand economy, forest products have a market value in excess of $2.6 billion annually and provide employment for twenty eight thousand people. Typically, management scenarios involve mechanical and/or chemical site preparation and planting of genetically improved seedlings, followed progressively by thinning and pruning, and then clear-fell harvesting within rotation intervals of approximately 20 to 30 years (McLaren, 1993). These same plantations are essential components affecting the scenic beauty of the New Zealand landscape, a key contributor to the quality of outdoor recreation experiences and to a growing tourism industry. Commercial forestry practices are encountering increased public objections from tourists, recreation visitors and more sensitive local residents, particularly in areas with high visibility.

The responsibilities of the commercial forest industry for visual/aesthetic value protection have been unclear in New Zealand's past. Nevertheless, forest managers have historically forgone scheduled harvests in some visually sensitive areas or have used techniques such as landscape screening and amenity planting in an effort to mitigate visual effects and maintain desirable public relations (Moore et *al.*, 1991;

Sissons and Conway, 1991). The more recent dedication of the forest industry in New Zealand to the improvement of visual environmental quality can be largely attributed as a response to the Resource Management Act (RMA) (Parliament of New Zealand, 1991). Thus, the provision of tools capable of effectively projecting and assessing the visual aesthetic consequences of alternative forestry practices is a challenge for researchers and essential for the modern forest manager to aid the development of appropriate policies and practices.

The RMA controls activities such as use, development or protection of the natural and physical resources of New Zealand and is based heavily on the investigation of the 'effects' of a proposed activity, rather than on prescribing which activities shall or shall not be allowed. It includes the ethnic philosophy of *Kaitiakitanga* - the exercise of guardianship of the land and, in relation to a resource, includes the ethic of stewardship based on the nature of the resource itself. In addition to consideration of 'effects', any mitigation efforts must be communicated clearly among the forest operator, regulatory authorities and other interested parties. The reasons for public objections to commercial forestry practices are diverse and complex, and visual impacts are a substantial contributor (Kilvert and Hartsough, 1993). Abrupt alterations of scenic environmental settings (Thompson and Weston, 1994) may pose direct threats to tourist and recreation industries, as well as residents environmental quality expectations. In addition, visual effects often precipitate public concern for other potential environmental and cultural impacts.

Generally the public are able to readily identify visual change in the landscape (Benson and Ullrich, 1981; Kilvert, 1995a; Kilvert, 1995b; Swaffield, 1994) and visual images are considered an excellent medium for communicating the effects of forestry operations to the public (Daniel and Boster, 1976; Daniel *et al.*, 1990; Orland, 1988; Orland, 1992). The idea of using images calibrated to known resource attributes to derive human values is not a new one. Malm *et al.* (1981) used image processing techniques to develop images of pollution plumes in the Grand Canyon, based on the output of numerical models of atmospheric dispersion that were used to derive human values for the impacts predicted on scenic resources in the canyon. The study established the effectiveness of computer techniques but used computing resources beyond the means of typical natural resource agencies. Specific applications with forestry relevance include Baker and Rabin's (1988) study of the visual effects of limb rust damage on national forest settings in northern Utah, the Orland *et al.* (1993) study of the impacts of insect damage and silvicultural responses on the scenery of the Dixie National Forest in southern Utah and Orland, Daniel and Haider's (1994) application to the visualisation of forest harvesting in Northern Ontario.

Although, activities such as forest harvesting have an obvious visual effect, it has been difficult to accurately calibrate visual landscapes to known levels of forest attribute or management activity. Pictures of forest harvesting have been used as illustrations of practice, rather than as one of the analytical tools in decision support. As part of an integrated study of forest harvesting values, a survey instrument was developed that used an extensive library of images to represent key variables related to anticipated forestry activities and the visual quality of the forest setting. The image set comprised computer scanned photography manipulated to reflect a range of attribute levels representing different management regimes and resulting changes over time. This report describes the procedures followed to develop the image sets, the validation procedures used for calibrating the imagery, details the perceptual survey techniques and presents the public response via an attitude survey to current and alternative forestry practices.

Method

Public Survey Design Issues

The verbal protocols common in pencil and paper surveys of public opinion can be generated automatically. Verbal phrases can be drawn from a look-up-table to fill the requirements of an experimental design and construct the survey instrument. This process works well with words

which serve then as abstractions of the kinds of conditions represented by the independent variables in the study. For instance, the words "Sixty metre buffer strip" stand for any combination of conditions that can buffer place A from place B by about sixty metres. The precise configuration or components need not be specified for the reader to have a mental image of what is intended.

This situation is quite different when using pictures that immediately make the mental image concrete. The same sixty metre buffer must be shown as vegetation, or not; as one species, many, or a mix; as a particular density or texture; on a realistic surface; and in the context of a surrounding matrix of other forest. The consequence is that, while a verbal phrase can be used repeatedly as a surrogate for a general concept of forest attributes, a picture implies a specific location and thus cannot stand as a surrogate for multiple situations. Moreover, seeing representations of the same resource attribute in different contexts brings into question the validity of how attributes are represented. Given our intention to develop visual protocols to address ranges of resource attributes, it was essential to address the constraints posed by images early in the design process.

The study required that a number of attributes would need to be represented visually but more significant was the interaction of those attributes in the visual display. Surrounding scenery can be shown as a separate issue but size of forest cutting operation cannot be separated from the forest type where it occurs, the shape and location of the cut, what is left as residual, or the stage of recovery of the cut. This distinction made it necessary to edit single images to match specifications from the experimental design so that the appropriate attributes could be seen concurrently.

Creating the Visual Instrument

The landscape images chosen to represent forest conditions in this study were taken at roadside locations in the Golden Downs and Rai Forests in the Nelson, NZ, area. Locations were chosen in concert with FRI staff and staff from the Fletcher Challenge company which manages many of the timber holdings in those areas. Table 1 illustrates

Table 1. *The design matrix of scene attributes*

Pre-harvest	Post-harvest Logging practice	2 years emerging visible rows Row orientation	8 years closed canopy	9/10 years pruned & thinned	20 years mature forest
Rai Bridge	hauler	contour	★	★	★
		vertical	★	★	★
Wai-iti	hauler	contour	★	★	★
		vertical	★	★	★
Inwoods	hauler	contour	★	★	★
		vertical	★	★	★
Kerr's Hill	hauler	contour	★	★	★
		vertical	★	★	★
	skidder	contour	★	★	★
		vertical	★	★	★
Norris Gully	buffer	contour	★	★	★
		vertical	★	★	★
	no buffer	contour	★	★	★
		vertical	★	★	★

the image set design with five initial images, four management treatments, and five time steps in scenario development. Implausible combinations were deliberately excluded from the set.

Developing the Set of Oblique Views

A first step to image creation was to create a library of source imagery. Since the roadside view was central to the study, the image library needed to be taken from a similar viewpoint. During the spring of 1994 more than three thousand photos of forest conditions were collected, around Golden Downs and Rai Forests near Nelson, and Whakarewarewa Forest at Rotorua.

The camera was a hand-held Nikon 8008 with autofocus and auto exposure. The majority of photos were taken at a 50 mm focal length setting. A moderate telephoto lens of 85 mm focal length was used at times for finer detail. The film was Kodak Ektachrome Elite, a 100 ASA semi-professional colour film with reasonable speed and good colour rendition. A polarising filter was used at all times and camera direction of view was held between 15 and 60 degrees of a line directly opposing sun bearing. These latter two measures were to maximise colour saturation. Three hundred images were selected from the entire set based on an appraisal of image quality as well as suitability for filling the experimental design requirements. These baseline images were transferred to Photo-CD format by Kodak.

A resolution of 768 x 512 pixels and 24-bit colour depth, was used for this project, a compromise between quality needed and the size and concomitant complexity of large image files. Adobe Photoshop™ software on Apple Macintosh computers was used for image manipulation. Orland (1988, 1993) has described the evolution of typical uses of these tools, the basic techniques, and issues of image validity and utility. All images underwent histogram equalisation to achieve the best consistent contrast and colouration throughout the image set as it was clear at the outset that the study design would necessitate considerable image editing and the use of an extensive source image library.

Implementation of SmartForest II

The New Zealand forest industry uses sophisticated forest management decision support systems (DSS) for growth and yield predictions, valuation and estate modelling. For our study, future forest stand conditions were predicted using STANDPAK (Whiteside, 1990), a DSS designed to model individual stand growth and yield while optimising silvicultural management alternatives. Forest harvesting plans were developed in consultation with forest managers. However, despite having carefully developed harvesting plans and projected future plantation conditions, to portray that information in a visual medium by constructing accurate data driven visualisations is a complex task.

Using a simple 3-D projection from GIS data does not adequate display the height of "layers" of trees on the landscape so that visibility can be verified, and the thickness of linear graphical elements is such that boundaries seen at oblique angles cannot be differentiated. In our study a great number of attributes need to be represented visually but equally significant was the interaction of those attributes in the visual display. For example, the size of a forest cutting operation cannot be separated from the forest type where it occurs, the shape and location of the cut, what is left as residual, or the stage of recovery of the cut. This distinction made it necessary to use a software system to create schematic analytic visualisations to match specifications from the experimental design and to verify that the appropriate attributes could be seen concurrently.

SmartForest II (Orland, 1994) is a landscape visualisation software package capable of displaying a schematic representation of tree density, size and homogeneity of stand composition in correct visual perspective and was designed to deal with planning forest landscapes at large scale but at the same time to be able to develop specific management strategies at a small, tree-by-tree scale. Real time display of viewsheds and the capability to move within the "scene" data space makes the tool eminently suitable for landscape planning applications. The software requires a Silicon Graphics or IBM RS6000 platform and is available

via the World Wide Web at http://imlab9.landarch.uiuc.edu/SF/SF.html.

For our application, each of the three component data sets necessary to drive SmartForest II simulations were provided by integration of the outputs from other software systems. A digital elevation model (DEM) to provide topographical data for creating the landform features was generated from elevation data stored as contour coverage's in a GIS, ARC/INFO® (ESRI, 1991). To derive a DEM involved stepping through a number of processes to convert the data to a grid of data in the USGS DEM format supported by both ARC/INFO® and SmartForest II. Problems were encountered in making the final transformation to the DEM, largely because of the differences in coordinate systems, units, and completeness of metadata in the NZ records versus the expectations of the ARC/INFO® software. Work-arounds were developed that involved manual editing of file headers to ensure an accurate DEM generation.

SmartForest II uses a Stand File to provide information about the location of stand boundaries to superimpose on the DEM, vector files containing data for each of the landscape management scenarios in our experimental design were developed using TerraSoft® GIS (PCI, 1996) before being translated to ARC/INFO® stand boundary coverages.

As SmartForest II utilises gridded data (where grid cells are assigned stand identifiers that determine the Tree List attribute data to be placed at that location) the stand boundary coverage's were translated into a grid format, for each age step to be visualised. Finally, the Tree List Files (which provide records of the vegetation to place in each stand cell location) were generated for each age class to be represented in the visualisations (0, 2, 8, 10, 20 years) by substitution with STANDPAK model data.

SmartForest II is effective at showing significant structural changes in canopy configuration, such as edge conditions created in harvesting, the effects of close range tree growth

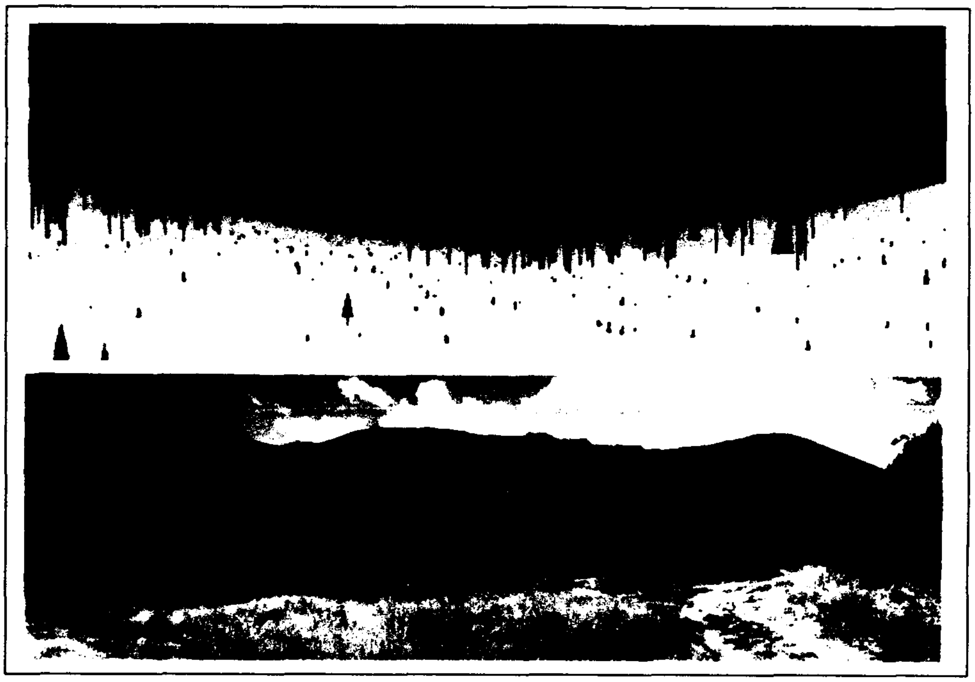


Figure 1: SmartForest II data driven analytical visualisation (upper) and Adobe Photoshop™ edited image (lower) for 1999 projection.

in buffer plantings on visibility of distant operations, and through use of false colour to show non-visible characteristics such as size class, age, species distribution and silvicultural history. SmartForest II visualisations can be stepped through a time sequence to generate analytical simulations for each viewpoint and scenario within the experimental design (Figure 1). Extensive use was made of the analytical simulations to guide image editing of the first sets of photorealistic visualisations. This approach provided a data driven and defensible linkage to the image product.

Creating Calibrated Images

Once the design issues were resolved, the assembly of the images was a somewhat mechanical process of taking image portions and combining them to fit the design specifications. Adobe Photoshop™ image editing software is the *de facto* standard for this task. Image editing processes are time-consuming and expensive and despite the extensive preparation work, it was difficult to achieve good fit between image parts. It was also intellectually taxing to synthesise the multiple concurrent demands of the study design into a *single* image. However, at this time the realism achievable by more directly data-driven visualisation tools is not good enough to support choices involving the appearance of scenic resources.

At three stages during the evolution of the image set, a process of intensive review and validation was undertaken.

One-on-one direct expert input

To further verify the shared understanding of forest conditions, the collaborators held an intensive workshop session to identify base and guide images for all scenes and to specify image editing directions for the image editors.

First-round review to verify image specifications and guidelines

After the first attempt to meet design specifications, all images were sent to FRI in draft form for review and feedback. Printed images were marked with instructions and returned to the image editors.

Further reviews to verify attribute scaling

As the image sets proceeded to completion it was critical to determine if they matched the desired attribute scaling, and if the forest conditions were represented accurately. Because of the distances between key participants, images were encoded as compressed JPG files and transmitted via ftp over the Internet.

The characteristics making the scenes particularly difficult both to design and construct as a visualisation were that the scenes represented many different forest areas in an extensive landscape setting such that stands would be at different viewer-object distances and orientations. All images detailed in the design (Table 1) were completed and successfully included in the attitude survey instrument.

Perceptual Survey

Visualisations of alternative forest plantation management scenarios were used in a systematic assessment of public perceptions of the visual consequences of each scenario. The perceptual assessment was approached in two formats: [1] a paired- comparison format in which overall visual effects of alternative management plans were represented across a full rotation for a single stand; and [2] a single-scene format in which the visual quality of individual views (each depicting only one stage of the progression from harvest to re-establishment to final mature growth) was rated in the context of a sampling of typical New Zealand plantation forest scenes. The rationale for the two procedures was that the paired comparison procedure provides the most sensitive assessment of perceived overall differences between the management options represented, while the single scene procedure more closely approximates the typical context in which a forest visitor might encounter the effects of forest management on the landscape.

The paired comparison survey was presented in an individual interview procedure applied to both New Zealand residents and to samples of foreign visitors. The single scene assessment was presented to groups of New Zealand residents and to one small group of foreign visitors.

Paired-comparison format

Visualisations for each management scenario for each represented site were laid out as individual colour prints arrayed on single A4 pages of a test "booklet" (photo-album). For each site-management scenario individual scenes were arrayed in a sequence from: [i] original condition (mature forest); [ii] immediately after harvesting; [iii] new forest (2 years after planting); [iv] young forest (8 years); [v] after thinning and pruning (10 years); and [vi] back to mature forest (20 years).

For four of the five forest sites represented (Rai Bridge, Wai-iti, Inwoods, and Kerr's Hill) the management plans compared differed only in whether the re-establishment of the forest following initial clear-cut harvest was accomplished by planting new trees in vertical rows (running up the slope) or in horizontal rows (following the contours of the slope), Figure 2. For the fifth site (Norris Gully) two pairs of scenarios were created, the first comparing vertical planting with and without a buffer of trees, Figure 3, (in this case larch) screening the harvested-planted area and the second comparing contour planting with and without the screening buffer.

The resulting six visualisation pairs were incorporated into the test booklet so that the two alternatives for each depicted forest site (vertical vs. contour planting, or buffer vs. no buffer) were displayed on facing pages of the booklet. Thus, participants were presented with pairs of visualisation pages, with each page presenting the six developmental steps (pre-harvest through cutting, planting and regrowth) for one of the management approaches (e.g., vertical or contour planting) simulated for a specific site (e.g., the Rai Bridge site). Comparisons always involved different management plans applied to a single site. Participants were not presented with any comparisons between the different forest sites. For each pair of visualised alternatives (facing pages), the participants were required to select the one alternative which in their judgement represented the best overall visual effects.

Single scene format

A selection of the individual scenes that composed the overall visualisations for each management alternative at each site were chosen for presentation to groups of New Zealand residents. For each site/management option, scenes were selected to represent forest conditions at pre-harvest, immediately after harvest, after two years, after 10 years and at full re- growth (20 years post-harvest). These individual visualisation scenes were rendered into colour slides, divided into two presentation sets and mixed with 30 additional colour slides depicting different scenes of typical forest plantation sites in various stages of harvest and re-growth.

Individual participant groups were shown one presentation set, so that each group saw between 3 and 5 "versions" (harvest/re-growth stages) of a given forest site, randomly mixed with 3 to 5 versions of each of the other visualisation sites, and 30 other scenes. The goal of this procedure was to better represent the context in which

Figure 2: Rai Bridge, age 8, contour (left) vs. vertical (right) planting schemes

Figure 3: *Norris Gully, age 2, vertical planting, buffer (left) vs. no buffer (right)*

forest management effects are typically experienced by forest visitors, e.g., many different sites are encountered, and each site is in one or another stage of development. Participants were required to rate each of the 50 scenes on a ten-point scale that extended from "very low scenic quality" to "very high scenic quality."

Participants and procedures

No attempt was made to achieve a formal "representative random sample" of either New Zealand residents or of foreign visitors for this perceptual assessment. Rather, a "convenience sample" procedure was employed, which has proven adequate in similar previous studies (e.g., Daniel & Boster, 1976; Malm, *et al.*, 1981). Candidates for the paired-comparison portions of the survey were intercepted by two interviewers at highly frequented locations in the region of the forest sites represented and asked to voluntarily participate. For the individual scene presentations to groups, an attempt was made to sample a cross-section of New Zealand resident groups that were a priori expected to have different perspective's and values regarding forest plantation management.

Paired-comparison sample

The paired-comparison assessment was conducted by individual intercept interviews carried out in the Nelson Region and in Christchurch on the South Island and in Rotorua on the North Island of New Zealand. Locations for the interviews were chosen to provide a target number of 500 participants, divided between New Zealand residents and foreign visitors. Interview sites were selected to maximise exposure to a diverse range of potential participants, where there was expected to be a relatively rapid turnover of people (such as a town center or a visitor information facility), and where people would be expected to have time available to complete the survey (such as the Interlslander Ferry and the airport). Given these criteria, a number of locations were selected.

The town centre, the Polytechnic, and the Airport locations were primarily aimed at the local resident population. The Interlslander Ferry was chosen as likely to provide a higher proportion of overseas visitors. Christchurch and Rotorua locations provided mostly New Zealand residents who lived outside the immediate area of the study. Other locations included the town center in Richmond, several towns near the study area, and nearby recreation areas.

The interviews were all conducted between 9 and 24 February, 1996. Interview times were selected to concur with observed peak time of occupancy which was typically around midday. The most successful areas in terms of number of participants were the Nelson Airport and the Interlslander Ferry, both due to the number of people passing through and to the relatively large amounts of time people spend at these locations. Also, people in these areas were typically seated, which apparently made them more willing to participate.

The intercept interviews were carried out with the use of the "booklet" (photo-album) of forest management visualisations described above. It was found that the most

effective way of interviewing the public was to approach them as they were waiting and ask them if they would like to participate. For their ease, the participants were given a clipboard and pen so they could take all the time required to study and complete the questionnaire. The questionnaire and answer sheet to the survey were designed to be as logical and as simple as possible so that it would apply to the greatest cross section of the public. For each visualisation-pair, participants were required to first choose the alternative which they judged as presenting the best overall visual quality, and then to allocate 100 "points" to indicate how much better the preferred alternative was. An allocation of 50/50 indicated no perceptible difference—the participant was simply guessing—and an allocation of 100/0 indicated that there was the maximum possible difference between the alternatives.

Some of the participants had difficulty understanding the printed instructions, so that it was necessary for the interviewer to take them verbally through the procedure resulting in more time being devoted to some respondents than to others. Each interviewer used two or more booklets, allowing the participation of more than one person simultaneously.

Individual-scene group sample

The group presentations complemented the intercept interviews by purposely targeting different sectors of the overall survey population (e.g., forest industry groups, community political groups, and environmental interest groups). Letters were sent out to local interest groups in the Nelson area, with the goal of contacting a wide spectrum of interests related to forest plantation management.

The group presentations were each of about three quarters of an hour in length. Each session comprised a brief introduction and instructions (postponing discussion of the objectives of the study) followed by presentation and rating of the 50 slides. Only after all the ratings had been completed, was there an explanation of how the visualisations were produced and the objectives of the study.

Demographic and other information

Each participant in both the paired-comparison and the single-scene presentation groups completed a brief questionnaire about themselves and the nature of their experiences and relationships to the New Zealand forests and landscape. Resident participants provided information about place of residence, ethnicity, frequency and contexts of visits to forest areas, family involvement with forest industry, and environmental group memberships. Non-residents provided information about country of residence, number of visits to New Zealand, purpose of present visit, and memberships in environmental interest groups.

In addition residents provided an estimate of the "contribution of the pine forests to the quality of the New Zealand landscape" by marking a line that extended from "extremely negative" to " extremely positive." Non-residents indicated three factors (scenery/landscape, food/accommodation, people, adventure activities, Maori culture, wildlife, horticulture, weather/climate) that "made the greatest positive contribution to quality of your visit" and also indicated their judgement of the contribution of the "pine forests to the quality of the New Zealand landscape" using the same line marking procedure as used by residents.

Results and Discussion

Visualisations

As part of the intent of this project to provide new forest management tools and to guide forest managers in satisfying multiple resource objectives it is important to evaluate the usefulness of the tools used in this case study in relation to operational forest management practices. Although the development of the visualisations was successful, the complex nature of the data and tools utilised for our visualisation process would require customisation, integration with individual databases and operator training before they could be used routinely in a production manner.

We recognised that issues of integration and accuracy of each of the diverse data sources would be critical to achieve photorealistic, defensible, data driven visualisations. Sev-

eral factors contributed to our success in providing visualisations of future forest conditions on a complex landform involving a number of individual forest stand components, namely: the accuracy of the spatial data, availability of well validated forest modelling systems and suitable landscape visualisation software.

SmartForest II software is a suitable tool for integration with routine forest management practices, however, all of the data sources necessary for its operation are not readily available for much of the New Zealand forest estate at present. Development of photorealistic imagery is time consuming, expensive and requires skilled operators. The key for forest managers however, is to realise what is possible with the increasing use of integrated GPS, digital photography, computing technologies and to simply start collecting the data.

We recognised a "spin-off" benefit during the image editing phase where it was necessary to generate a consensus view among a panel as to the representation of conditions on the ground. We found that this view may, in fact, be more valuable information than a precise biological description. While there is an obvious need to supply the management process with better information, the collective judgment of experienced managers is also a valuable source of data. A possible by-product of such collaborative reviews may be a better grasp of the salient issues and a better shared understanding within a management group. This speculation is untested, but based on our interpretations of observing other review processes.

Perceptual Survey

The primary results of interest were the expressed preferences among the visualised alternative management scenarios in the paired comparison interviews and the scenic quality ratings provided by the groups in the single-scene assessment. In that context, comparisons were made in the assessments provided by residents and non-residents, and among the various sample locations and interest groups represented. Results for the paired comparison and the single scene portions of the study are reported separately below.

Paired-comparison results

As indicated in Figure 4, participants did not exhibit differential preferences for the visualisations of vertical planting as compared to contour planting methods. The Rai Bridge, Wai-iti, Inwoods and Kerr's Hill sites all produced average point allocations for the Vertical option that were very near (and not significantly different from) 50/50. With regard to the comparisons of buffer vs. no buffer (the Norris Gully site), the visualisations that retained the buffer of larch trees screening the harvested area was consistently and substantially preferred for both the vertical and the contour planting options represented, with point allocations averaging 80/20 in favour of buffers in both cases.

Figure 4 also reveals that there was very little (not significant) variation among the responses recorded by the participants intercepted at the various interview locations. Similar comparisons also failed to find significant differences between residents and foreign visitors. Figure 5 compares the responses of participants indicating different relationships to the forest landscape—those indicating direct personal or family involvement in forest industry, those indicating membership in environmental interest groups and others. The same pattern of lack of preferences between vertical and contour planting schemes, and substantial preference for buffers are repeated, with no discernible differences among these interest-defined groups.

Single-scene results

Ratings for each of the single scenes were averaged across all group respondents. Results for two representative sites are graphed in Figure 6 (comparing contour and vertical planting schemes for the Rai Bridge site) and Figure 7 (comparing buffer and no buffer scenes for the vertical planting scheme at Norris Gully.

While the paired comparison results indicated no difference in preferences for the overall visual effects of vertical and contour planting approaches, there is some indication that contour planting was judged to produce higher scenic quality at the two points (age 2 and age 10 -post thinning and pruning) where the two planting patterns would be the most visually conspicuous. The pattern of ratings

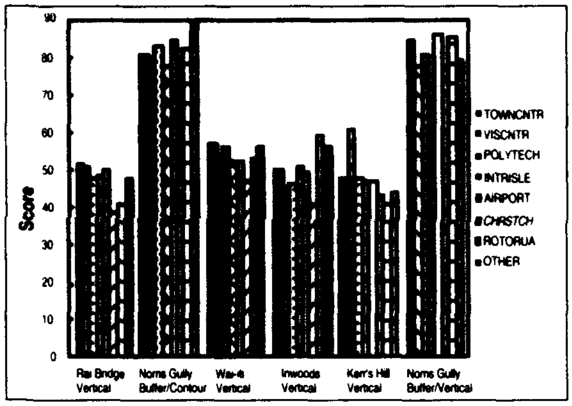


Figure 4: *Paired comparison scores by location and site*

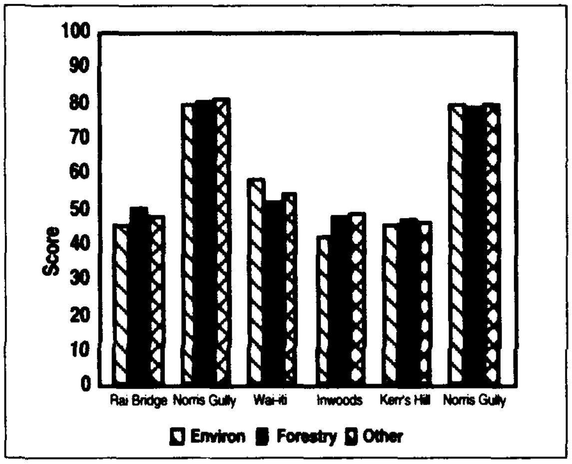


Figure 5: *Scores by relation to forest landscape groupings and site*

for the individual scenes depicting buffer and no buffer conditions is consistent with the overall preferences expressed in the paired-comparison assessment where buffered scenes were consistently preferred.

Questionnaire results

Table 2 presents some of the more important results from the brief demographic-forest experience questionnaire. Of particular interest is a comparison of the importance ascribed to the "pine forests" contribution to the quality of the New Zealand landscape by residents and visitors. These data were derived by measuring the location of the marks on the line provided on the response form, and then transforming that measurement to a scale that extended from 0 at the lowest end to 10 at the highest. As the table reveals, Nelson area residents (closest to the study area) indicated the lowest (slightly negative) opinions, followed by other New Zealand residents (slightly positive). The foreign visitors tended to ascribe very similar, and substantially higher values.

Non-resident visitors were also asked to indicate what factors contributed most to their

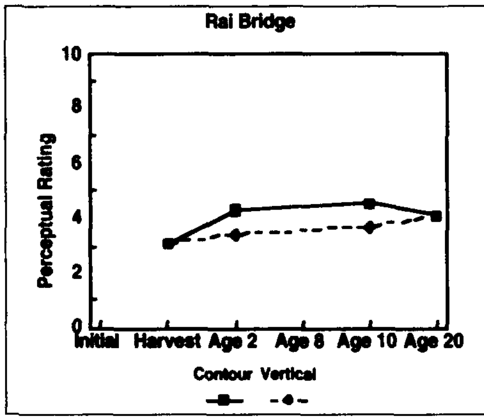


Figure 6: *Perceptual rating for contour versus vertical planting schemes*

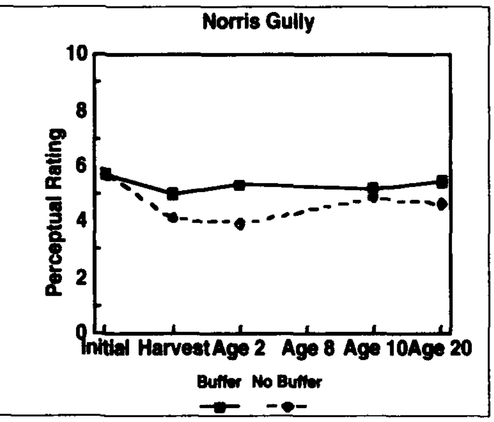


Figure 7: *Perceptual ratings for buffer versus no-buffer for vertical planting schemes*

Table 2: Contribution of pine forest to the New Zealand landscape (score), averaged by resident type and resident origin

Resident Type	Resident Origin	Number of Respondents	Score
Residents	Nelson and District	185	4.75
	Other South Island	85	6.10
	North Island	98	5.94
Non-Residents	Europe	81	7.40
	North America	32	7.60
	South America	12	7.50
	Asia & Others	9	7.90
		502	

enjoyment of their visit. Based on the number of positive responses recorded, scenery was the most important positive factor (109 responses), followed by people (76 responses) and weather/climate (62 responses).

Acknowledgments

This project was supported in part by a USDA Forest Service Co-operative Contract 28-C5-871, by the Foundation of Research Science and Technology and by Fletcher Challenge Forests.

We would particularly like to thank Nigel Brabyn and Nick Roberts of Fletcher Challenge Forests; Emma Haines of the University of Hertfordshire; Andrew Dunningham and Graham West of the New Zealand Forest Research Institute; William White, Ross Pywell, Jeanine Paschke of the US Forest Service FHTET; Larissa Larsen, Ai-lin Shu, Ryan Palm, Kittipong Mungnirun, Kenneth Schalk, and Kaiyu Pan of the Imaging Systems Laboratory, University of Illinois, for their contributions to the project.

References

Baker, F.A. and Rabin, D. (1988) Using computer graphics to assess the visual impact of limb rust in Ponderosa pine. *Utah Science*, Utah Agricultural Experiment Station, Logan: UT. 49(4):98-102

Benson, R. E. and Ullrich, J. R. (1981) Visual impacts of forest management activities: Findings on public preferences. *USDA Forest Service Research Paper INT-262 Ogden, Utah.*

Daniel, T. C. and Boster, R. S. (1976) Measuring landscape aesthetics: The scenic beauty estimation method. *USDA Forest Service Research Paper 167, Ft. Collins, Colorado.*

Daniel, T. C., Orland, B., Hetherington, J. and LaFontaine, J. (1990) Model-driven visual simulations. *In:* J. H. Gramman (Ed.) Proceedings of the third symposium on social science and natural resource management, College Station, Texas, May 16-19.

ESRI. (1991) ARC/INFO®. *Environmental Systems Research Institute, Redlands, California, USA.*

Kilvert, S. and Hartsough, B. (1993) Visual impacts of forest operations: Measuring concern in New Zealand. *LIRO Report 18 (9).*

Kilvert, S. (1995a) Forest landscape change: What do the public think? *In:* Proceedings of the LIRO Harvest Planning Conference, Nelson, New Zealand.

Kilvert, S. (1995b) Kawau Island Historic Reserve: A public perception of logging and landscape change. *LIRO Report* 20 (8).

Malm, W., Kelley, K., Molenar, J., & Daniel, T.C. (1981) Human Perception of Visual Air Quality (Uniform Haze). *Atmospheric Environment*, 15(10/11), pp1875-1890.

McLaren, J. P. (1993) Radiata pine growers manual. *New Zealand Forest Research Institute Bulletin No. 84.*

Moore, J., Orwin, J., Weston, J. (1991) On the edge management options for plantation edges. *Ministry of Forestry.*

Orland, B. (1988) (2nd ed., 1992) Aesthetic preference for rural landscapes: Some resident and visitor differences. Ch. 5. *In*: J. Nasar (*Ed.*) The visual quality of the environment: Theory, research and applications. *Cambridge*. pp. 364-367.

Orland, B. (1992) (*Guest Ed.*) Data visualization in environmental management. *Special Issue Landscape and Urban Planning 21*.

Orland, B. (1993) Synthetic landscapes: a review of video-imaging applications in environmental perception research, planning, and designs. *In*: D. Stokols and R. Marans (Eds.) Environmental Simulation: Research and Policy Issues. *new York: Plenum*.

Orland, B. (1994) SmartForest — A 3-D interactive forest visualization environment. *In*: Proceedings, Decision Support 2001 — Resource Technology 94, American Society for Photogrammetry and Remote Sensing, Washington, D.C. pp. 181-190.

Orland, B., Daniel, T.C., Hetherington, J. and Paschke, J.L. (1993) *Visualization of Forest Management Issues on the Dixie National Forest*. Final report. US Forest Service Forest Pest Management — Region 4 and Methods Application Group (FPM-MAG). 60pp.

Orland, B., Daniel, T. C. and Haider, W. (1994) Calibrated Images: Landscape Visualisations to Meet Rigorous Experimental Design Specification. *In*: Proceedings, Decision Support 2001 - Resource Technology 94, ASPRS, Washington, D.C. pp. 919-926

Parliament of New Zealand. (1991) Resource Management Act 1991.

PCI. (1996) TerraSoft® *PCI, Ontario, Canada*.

Sissons, D. and Conway, M. (1991) Golden Downs Forest Clark Valley - A proposal for reducing the visual impact of timber harvesting. *Sissons and Conway Limited. 23 Crop Place, Richmond, Nelson*.

Swaffield, S. R. (1994) Attitudes towards trees: a case study in the New Zealand high country. *NZ Forestry February, 1994*. pp. 25-30.

Thompson, S. and Weston, J. (1994) South Island high country forest design. *Ministry of Forestry*.

Whiteside, I.D. (1990) STANDPAK Stand modelling system for radiata pine. *In*: Proceedings IUFRO Symposia "New approaches to spacing and thinning in plantation forestry" April 10-14, 1989, Rotorua, New Zealand. *Ed*: R. N. James and G. L. Tarlton

Visualisation Techniques for Collaborative GIS Browsers

Neville Churcher
Department of Computer Science,
University of Canterbury,
Private Bag 4800, Christchurch
neville@cosc.canterbury.ac.nz
voice/voicemail: (03) 3642353; fax: (03) 3642569

Parames Prachuabmoh and Clare Churcher
Centre for Computing and Biometrics,
Lincoln University,
P.O. Box 84, Lincoln
{prachuaj,churchec}@lincoln.ac.nz

Presented at the second annual conference of GeoComputation '97 & SIRC '97, University of Otago, New Zealand, 26-29 August 1997

Abstract

Visual information overload is a serious problem for users of geographical information systems (GIS), or other applications with complex displays, where the requirements of access to both local detail and wider context conflict. This problem is compounded for users of real-time groupware applications by the need to maintain awareness information about other users and their actions. In this paper, we describe our use of fisheye views to assist with visual information overload management in GROUPARC, a lightweight real-time groupware application for browsing and annotating GIS data.

1 Introduction

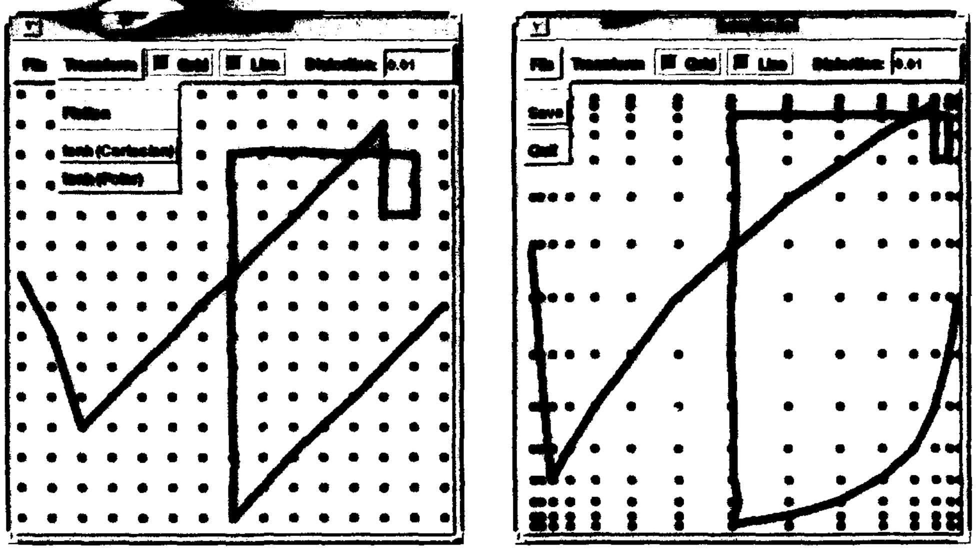
Our capacity for assimilating complex visual displays, such as GIS data, is limited. The phenomenon of visual information overload occurs when this capacity is exceeded, typically resulting in confusion, oversight and errors of interpretation.

The ability to focus on regions of interest in detail, while retaining awareness of context, is necessary if users are to visualise and comprehend complex graphical information effectively. It is important for users of GIS to be able to examine not only local feature detail (e.g. utility access points on some land parcels) but also to be aware of related but spatially separated features (e.g. high voltage network).

Conventional approaches to this problem include scrolling, zooming and split windows. However, each has its faults (see e.g. Churcher 1995a) both in terms of cognitive load for the user and clutter of the precious display real estate.

The terms "fisheye view", "distortion-oriented presentation" and "non-linear magnification" are among those used to describe visualisation techniques where the displayed image is transformed in some non-uniform manner. Since Furnas's (1986) introduction of the concept, there has been much interest in these techniques as a means of improving the usability of complex graphical displays (for a bibliography see Keahey 1997). While cartographers have made effective use of exotic projections for some time, the extension to dynamic interactive interfaces is more recent (Sarkar & Brown 1992, Sarkar & Brown 1994, Churcher 1995b).

The central idea is to emphasize "relevant" regions of the display, and de-emphasize less relevant areas, without loss of context. This is achieved by transformations which distort the distances between features while preserving connectivity and topological relationships. Figures 1 and 2 show some examples produced from a teaching tool we have developed. An important concept is the focus—a region where interest is concentrated—and distance from the focus is part of the measure of relevance or importance. Fisheye transformations are discussed further in section 2.



(a) No transformation (flat)

(b) Transformed (equation 2)

Figure 1: Experimenting with fisheye transformations

The problem of visual information overload is particularly important for Computer Supported Collaborative Work (CSCW) applications—also referred to as "groupware". Baecker(1993) provides a good overview of groupware. Simple examples such as drawing tools are becoming commonplace in the commercial environment but there are many challenges associated with extending the concept to include "serious" applications such as GIS.

The collaborative GIS browser GROUPARC (Churcher & Churcher 1996b, Churcher & Churcher 1996a) is an example of a GIS groupware application. It is a flexible lightweight tool enabling users located anywhere on the internet to share, examine, discuss, annotate and visualise GIS data in real time using using a What You See Is What I See (WYSIWIS) model. It might be used in situations as diverse as a classroom exercise or a geographer in the field debating planning options with colleagues in another country.

Users of CSCW GIS applications must not only contend with the problems discussed above but also with the processing of additional information associated with awareness of other participants in the conference.

Maintaining each participant's awareness of the presence, location, intentions and actions of others is an essential element of successful groupware and innovative techniques are being developed to address the issue (e.g. Greenberg, Gutwin & Cockburn 1996). GROUPARC's approach is discussed in detail elsewhere (Churcher & Churcher 1996b, Churcher & Churcher 1996a) and in subsequent sections.

There is currently much interest in developing CSCW GIS applications (Armstrong 1993, Armstrong 1994, Faber et al. 1994, NCG 1995, Jones et al. 1997). We envisage the gradual introduction of both CSCW capabilities and distortion-oriented presentation techniques into mainstream commercial GIS products over the next few years. Each is important in its own right.

Our current research concentrates on lightweight browsers rather than fully-featured GIS systems. There are a number of specific differences. Firstly, GROUPARC allows users to work with GIS data without requiring them to have the same GIS software—or any GIS at all! Consequently lightweight tools such as GROUPARC offer an alternative to simply waiting for vendors to embrace standards. It is envisaged that users will still turn to a fully-fea-

tured GIS for resource intensive tasks such as complex spatial queries and topological analysis. Lightweight tools offer extensive opportunities for extension and customisation in order to find the most appropriate solution (e.g. choice of transformation function) for each problem. Finally, portability across platforms (hardware, communications and operating system) is straightforward.

The remainder of the paper is structured as follows. In the next section we discuss fisheye views further and introduce the particular forms of fisheye view that we have incorporated into our latest version of GROUPARC. Section 3 contains a brief summary of GROUPARC's GIS and CSCW features and indicates how fisheye techniques have been incorporated naturally. In section 4 we discuss some of the approaches we have explored, present results showing some of the techniques we have implemented, and comment on the relative suitability of each for GIS applications. Finally, some conclusions and indications of the future directions of our research are presented in section 5.

2 Fisheye views

An essential ingredient in any fisheye interface is a spatial transformation function, G, which maps a "flat" coordinate value, x, onto the corresponding transformed value, x'. The derivative G' is the corresponding magnification function. The main transformation function used in our current work is based on the tanh function (Keahey & Robertson 1996) which has the general form shown in equation 1 for one dimensional coordinates.

$$x' = \tanh(\beta x) \qquad (1)$$

where β is a scalar parameter.

The tanh transformation maps coordinate values x in the range $[-\infty, \infty]$ onto corresponding values x' in the range $[-1, 1]$. It is very similar in its effect to that of the function

$$G(\hat{x}) = \frac{(d+1)\hat{x}}{d\hat{x}+1}$$

made popular by Sarkar & Brown(1992, 1994) but is easier to work with in practice.

For GIS, we require the transformation to map the flat display region onto itself, in order to minimise jarring visual effects. In particular, the focal point, and points on the boundary, should be invariant while other points should all move away from the focus towards the boundary.

For our purposes it is also important to be able to move the focus to any point within the display to enable users to see most clearly the portion of the display under most active discussion. In practice, users will move the focus precisely to attract attention to a specific area. If we consider values of x in the range $[0, x_{max}]$ with the focus, x_f in the same range then we should replace the transformation of equation 1 with

$$x' = \begin{cases} \tanh(\beta\,(x - x_f))(x_{max} - x_f) + x_f & (x > x_f) \\ \tanh(\beta\,(x - x_f))x_f + x_f & (x \le x_f) \end{cases} \qquad (2)$$

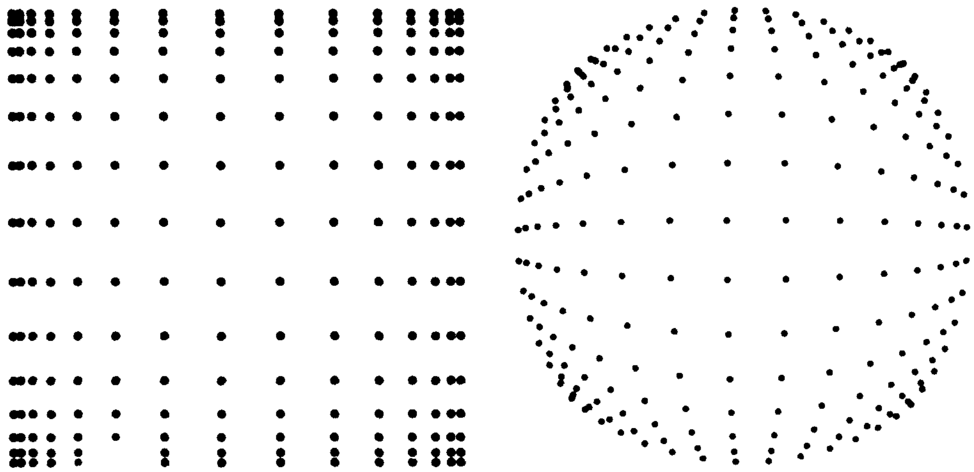
Extension to 2-dimensions, essential for any GIS application, is generally achieved using an orthogonal (Cartesian) or a polar (radial) approach. Further description of these and other approaches is available elsewhere (Keahey & Robertson 1996, Keahey 1997, Leung & Apperley 1994). Figure 1 shows a simple application we have developed to experiment with the effects of varying the parameters and functional form of G and some sample output appears in figure 2.

In the Cartesian form, the 1-dimensional transformation of equation 2 is applied independently to the x and y coordinates. The effect of this transformation is visible in figure 2(a). Under this transformation horizontal/vertical lines remain horizontal/vertical but, in general, angles are not preserved (as can be seen in figure 1). It is possible to apply transformations of different powers to each dimension (i.e. $\beta_x \neq \beta_y$) though we have not found it useful to do so.

In the polar form, the distances involved are not along the x or y coordinate axes but rather along the vector $\hat{\mathbf{p}} = \mathbf{p} - \mathbf{f}$ from the point $\mathbf{p} \equiv (x, y)$ to the focus $\mathbf{p}_f \equiv (x_f, y_f)$. The radial component of $\hat{\mathbf{p}}$ is then given by

$r = \sqrt{\hat{p}_x^{\,2} + \hat{p}_y^{\,2}}$ and the polar counterpart to equation 2 is

$$\mathbf{p}' = \frac{\tanh(\beta\, r)}{r}\hat{\mathbf{p}} + \mathbf{f} \qquad (3)$$



(a) Car[illegible] β=0.01)　　　(b) Polar (β=0.01)

Figure 2. Comparison of Cartesian and polar tanh transformations (focus at centre of display).

Figure 2(b) shows the polar transformation of equation 3. The effect is familiar as it resembles that of the ultra-wide angle "fisheye" lens used in photography. Although this transformation bends horizontal and vertical lines it does preserve angles more closely. Though we haveyet to perform controlled user studies, our experience to date supports Sarkar & Brown's (1992) observation that users preferred the polar version of their transformation for geographical data.

3 GROUPARC

GROUPARC was initially developed to explore the potential of lightweight CSCW browsers for GIS applications. It is written in Tcl (Ousterhout 1994), runs on Unix, Macintosh and Windows platforms and uses GroupKit (Roseman & Greenberg 1992, Roseman & Greenberg 1996), a toolkit for building real-time groupware applications (called conferences). When GROUPARC is running, GroupKit manages the registration of conference participants (who may enter or leave at any time) and communication between the GROUPARC replicas on individual participant's workstations. Typically, users will be participating in several additional conferences—such as editors and drawing tools.

GROUPARC users load one or more coverages (thematic layers) and then explore and annotate them with text and sketches during the course of a discussion. The coverage stacking order is reflected by shading and may be modified by users to handle co-located features.

These figures show several aspects of typical GROUPARC use scenarios. User-selected characteristic colours are used to distinguish individuals. Multi-user scrollbars, consisting of an ordinary scrollbar plus an indicator showing the relative positions of other users, are visible and show that there are currently three participants whose viewing regions may overlap (figure 3) or diverge(figure 4(a)). Telepointers, which show remote users' cursors as blobs of their characteristic colours, are a further awareness indicator. A telepointer is visible in figure 3 near the check mark beside the text "Soil analysis".

Figures 4(a)-5(d) show a single coverage of data about roads in part of Christchurch. The GROUPARC image window (figure 4(a)) shows a GIF image which has been annotated as the three conference participants acquaint themselves with the location of the region to be discussed.

A particular arc has been selected (thick line) as the response to a query ("which arc has $recno = 613?"). The

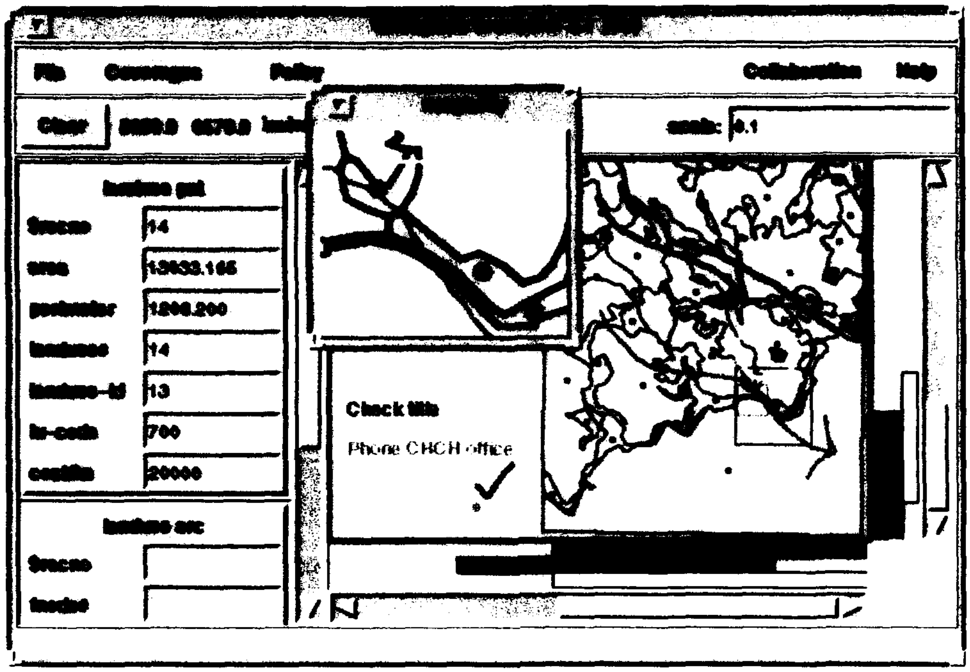


Figure 3: Typical GROUPARC session.

arc immediately to the right has been highlighted, as the user's cursor (not shown) is currently over it, and the corresponding attribute data are shown. The text annotation "My house" and the sketched circle have been added by other users.

4 Implementation & experience

Experience with GROUPARC has indicated clearly that loss of context is a problem as users focus on local detail. In this section we illustrate some of our approaches to date.

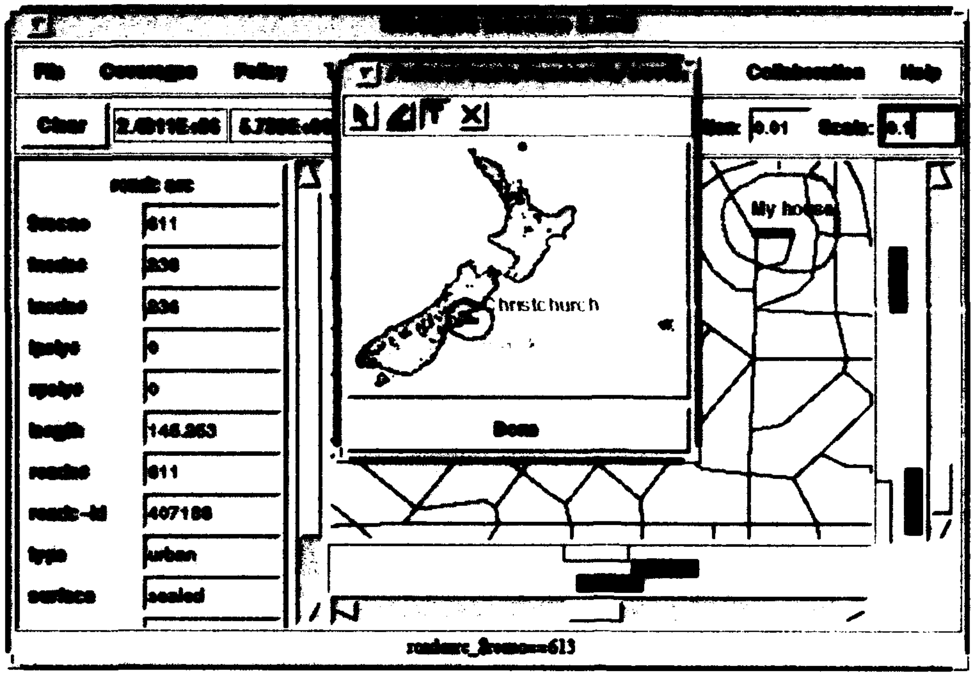
The simplest solution we implemented (figure 3) provides each participant with a floating window containing a uniformly magnified view of part of the main display. The main GROUPARC window contains rectangles (coloured to represent the corresponding users) which show the regions each user sees in the magnified window. These may be dragged around – typically to enable users to align their high-detail regions.

This technique is similar to that of the offset lens (Greenberg et al. 1996) and is particularly effective where the data is relatively uniformly detailed. In such cases fisheye transformations move many peripheral features to nearly identical locations leading to densely cluttered regions.

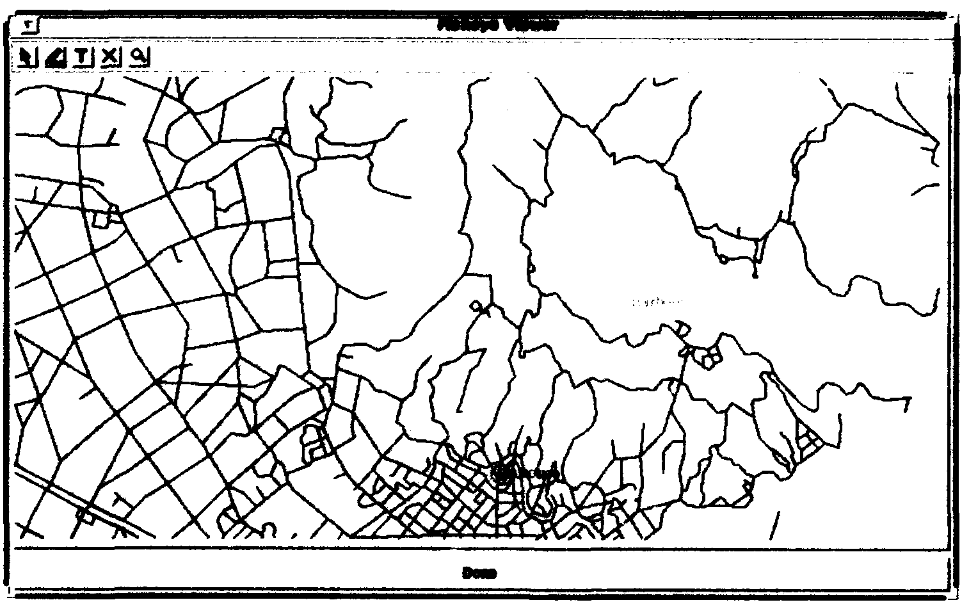
Figure 4(b) shows the entire coverage fitted into a window ready for transformation. The position of the focus is indicated by the magnifying glass at the centre of the figure.

Figure 5(a) shows the effect of the transformation of equation 2 with the focus remaining at the centre. All features have moved away from the focus, as expected from figures 1 and 2, and the arcs (including sketch annotations) have been distorted. The text annotations have also moved but, for clarity, their size has not been changed.

Figure 5(c) shows the effect of moving the focus close to "My house". Applying the transformation of equation 3 with the focus at the centre and "My house" produces the displays of figures 5(b) and 5(d) respectively.



(a) Main GROUPARC display



(b) Undistorted
Figure 4 Christchurch roads coverage

We have not yet performed comprehensive user evaluations of our fisheye additions. However, anecdotal evidence from our colleagues and students suggests common themes. Firstly, the system has proved easy to learn and use and we believe a single user-controlled parameter is more natural than the 5 used in Sarkar & Brown's (1992) system. Polar transformations seem intuitively more appealing and users report greater difficulty judging distances and orientations in the Cartesian form. The addition of grid lines as a background cover might help. The polar transformation also seems to be preferred where the focus is near the edge of the display, where the Cartesian form tends to give a crush of features. The simple floating zoom window has proved surprisingly popular. It also avoids the perception that the space between features is being magnified.

Given that Tcl is an interpreted language, the efficiency of the transformation is satisfactory —typically 7 seconds for the roads cover on an 85MHz SPARCstation 5 — and users have not commented adversely about response times. The roads cover consists of 791 arcs composed of 2390 points. Distortion is achieved by repositioning the points so the density of points used in digitising can affect the smoothness of the result. Our experience with other applications suggests that an order of magnitude improvement may be obtained by implementing critical functions in C.

We are currently exploring two major directions. Firstly, our experiences suggest that hybrid transformation functions are likely to be superior and we are currently developing these. Hybrid transformations uniformly magnify points within a specified region centred on the focus and non-linearly transform points outside this region with a smooth transition at the boundary. Some work on such functions has recently been reported by Keahey & Robertson (1996).

The second direction represents more of a step towards Furnas's (1986) original concept of transforming features according to their degree of interest (DOI), rather than purely spatial location. A feature's DOI includes contributions from its a priori interest (API) and its distance (D) and in the 1-dimensional case has the form

$$DOI(x \mid x_f) = API(x) - D(x, x_f). \quad (4)$$

Each feature's API depends primarily on its non-spatial attributes and is independent of the location of the focus. In the case of GIS applications, factors contributing to the API might include the coverage (e.g. "roads are more relevant than rivers"), attribute values (e.g. "sealed roads are more relevant than metalled roads") or coarse spatial properties (e.g. roads in our province are more relevant than those in neighbouring provinces).

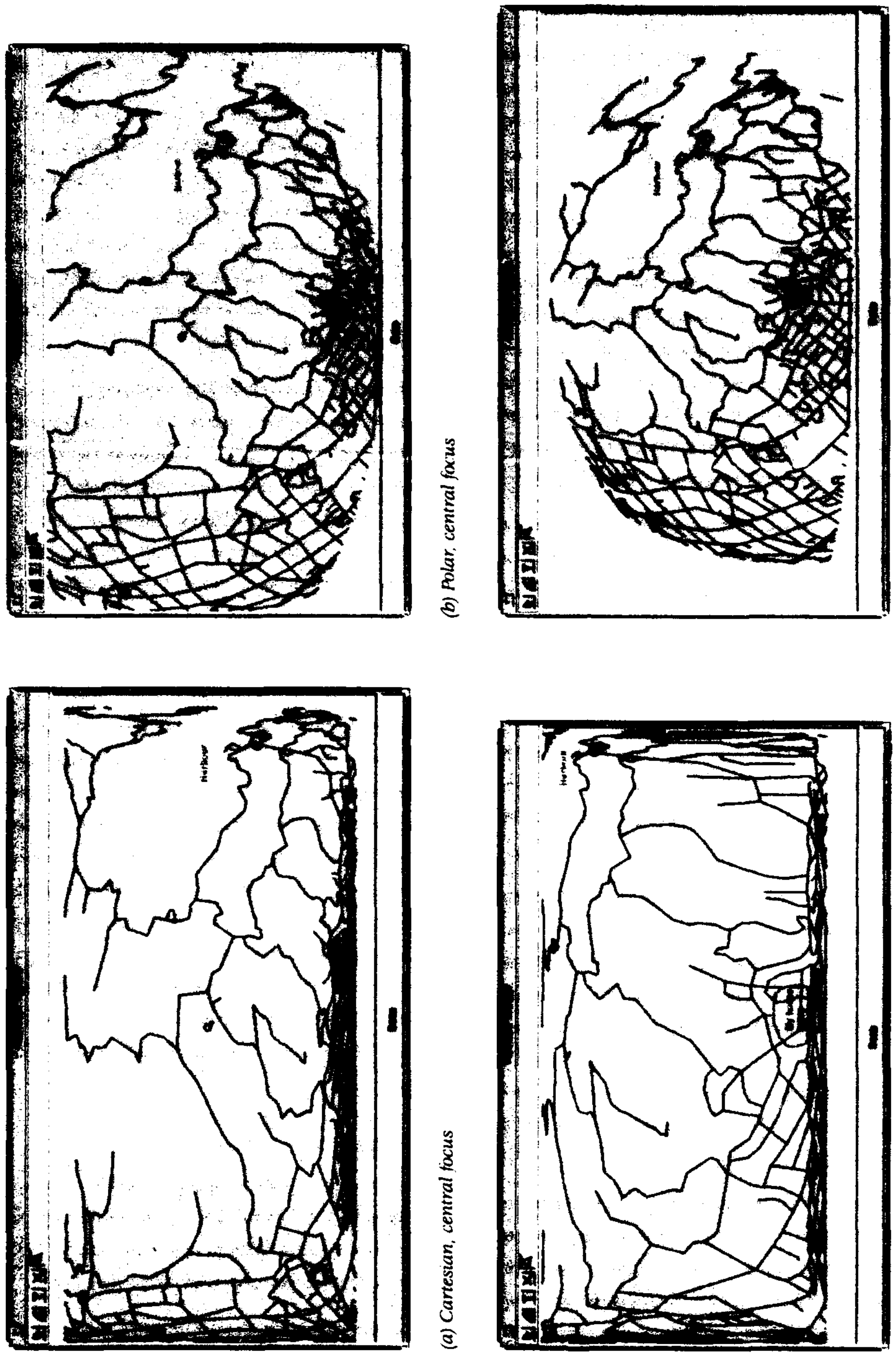
The distance is measured from the feature to the focus and may include contributions from "conceptual distance" as well as pure spatial distance. For example, the distance between two urban locations may be the straight line distance between them weighted by the "Manhattan" distance between them and the number of traffic lights along the route. The (focus-independent) API and (focus dependent) distance can combine in such a way that the overall DOI for a "very interesting" feature far away is similar to that of a "less interesting" feature nearby.

The display is then presented in such a way that higher prominence is given to the most relevant (i.e. largest DOI) features at the expense of less relevant (lower DOI) ones. This approach suggests a solution to the problem of dense regions produced by transformations. As its DOI decreases, a feature becomes progressively de-emphasized and ultimately omitted from the display when its DOI becomes less than a user-selected threshold value. For example, labels may cease to be displayed when their font size becomes too small to read, extended features may be represented by points and colour may be replaced by monochrome.

5 Conclusions

We are encouraged by the success of our addition of fisheye capabilities to GROUPARC. They have proved useful in helping users visualise GIS data not only in GROUPARC sessions with others but also in the single user case.

Our current efforts are directed towards adding hybrid



(a) Cartesian, central focus

(b) Polar, central focus

(c) Cartesian, focus at 'my house'

(d) Polar, focus at 'my house'

Figure 5: Fisheye views of the Christchurch roads coverage

transformation functions and developing an interface to support user-selected API functions and DOI thresholds. The API will be specified by selecting from the available coverages and placing constraints on attribute values using the existing query functionality. Users will then have a natural, problem-related means of achieving a high degree of control over the transformation details. We will then optimize for performance by implementing the transformation functions in C before proceeding with controlled user trials.

We also intend to investigate the potential uses of multiple focus points, one per conference participant, which allow several regions of interest to be examined in greater detail simultaneously.

6 Acknowledgements

We would like to thank Ryan Clements for providing us with access to data and two anonymous referees for their helpful comments.

References

Armstrong, M. (1993), 'Perspectives on the development of group decision support systems for locational problem solving', Geographical Systems 1(1), 69-81.

Armstrong, M. (1994), 'Requirements for the development of GIS-based group decision support systems', Journal of the American Society for Information Science 45(9), 669-677.

Baecker, R. (1993), Readings in Groupware and Computer Supported Cooperative Work: As sisting Human-Human Collaboration, Morgan Kaufman.

Churcher, C. & Churcher, N. (1996a), Real-time conferencing in GIS, Technical report, Depart ment of Computer Science, University of Canterbury.

Churcher, N. (1995a), Applications of distortion-oriented presentation techniques in GIS, in P. Firns & N. Sutherland, eds, 'New Zealand Conference on Geographical Information Systems and Spatial Information System Research', AURISA/SIRC, SIRC University of Otago, Massey University, pp. 323-336.

Churcher, N. (1995b), 'Photi—a fisheye view of bubbles', Information & Software Technology 37(1), 31-37.

Churcher, N. & Churcher, C. (1996b), GROUPARC—A Collaborative Approach to GIS, in R. Pascoe, ed., 'Proc. 8th Annual Colloquium of the Spatial Information Research Centre', University of Otago, pp. 156-163.

Faber, B., Knutson, J., Watts, R., Wallace, W., Hautaluoma, J. & Wallace, L. (1994), A *groupware-enabled GIS*, in M. Heit, H. D. Parker & A. Shortreid, eds, *'GIS'94: GIS Applications in Natural Resources 2'*, Vanvouver, B.C., pp. 3-13.

Furnas, G. (1986), Generalised fisheye views, in 'Proc ACM SIGCHI '86 Conference on Human Factors in Computing Systems', pp. 16-23.

Greenberg, S., Gutwin, C. & Cockburn, A. (1996), Using distortion-oriented displays to support workspace awareness, in 'Proceedings of the 1996 British Computer Society Conference on Human-Computer Interaction. Imperial College, London. 20-23 August', Cambridge University Press.

Jones, R., Copas, C. & Edwards, E. (1997), 'GIS suppoert for *distributed group-work in regional planning*', Int. J. Geographical Information Science 11(1), 53-71.

Keahey, T.A. (1997), 'Nonlinear magnification home page', http://www.cs.indiana. edu/hyplan/tkeahey/research/nlm/nlm.html Computer Science Depart ment, University of Indiana.

Keahey, T.A. & Robertson, E. L. (1996), Techniques for nonlinear magnification transforma tions, in 'Proceedings of the IEEE Symposium on Information Visualization', pp. 38-45.

Leung, Y. K. & Apperley, M. D. (1994), 'A review and taxonomy of distortion-oriented presen tation techniques', ACM Transactions on Computer-Human Interaction 1(2), 126-160.

NCG (1995), 'NCGIA initiative 17: Collaborative spatial

decision making', http://www.ncgia.ucsb.edu/research/i17/I-17_home.html

Ousterhout, J. K. (1994), Tcl and the Tk Toolkit, Addison-Wesley, Reading, Massachusetts.

Roseman, M. & Greenberg, S. (1992), Groupkit:A groupware toolkit for building real-time conferencing applications, in 'Proc. CSCW '92'.

Roseman, M. & Greenberg, S. (1996), 'Building real-time groupware with GroupKit, a group ware toolkit', ACM Trans. Computer-Human Interaction 3(1), 66-106.

Sarkar, M. & Brown, M. (1994), 'Graphical fisheye views', Communications of the ACM 37(12), 73-84.

Sarkar, M. & Brown, M. H. (1992), Graphical fisheye views of graphs, in 'Proc. ACM SIGCHI '92', pp. 83-91.

Extraction of Beach Landforms from DEMs using a Coastal Management Expert System

Tony Moore, Kevin Morris, Grahame Blackwell[*] and Steve Gibson

Plymouth Marine Laboratory
Prospect Place, West Hoe,
Plymouth, Devon,
PL1 3DH
UK
Tel: +44 (0) 1752 633408
Fax: +44 (0) 1752 633101
abmo@unixmail.nerc-pml.ac.uk

[*]School of Computing,
University of Plymouth,
Drake Circus, Plymouth,
Devon, PL4 8AA UK
Tel: +44 (0) 1752 232541
Fax: +44 (0) 1752 232540

Presented at the second annual conference of GeoComputation '97 & SIRC '97, University of Otago, New Zealand, 26-29 August 1997

Abstract

This paper describes the use of a prototype coastal management expert system, to facilitate the extraction of a salient element of coastal landforms from a DEM derived from stereo aerial photography. The system, COAMES (COAstal Management Expert System), is currently under development at Plymouth Marine Laboratory and the University of Plymouth. The sub-landform to be extracted is identified and isolated through use of "intelligent" ground control points stored within COAMES' object-oriented data structure, in conjunction with geomorphological rules and functions embedded in COAMES' hierarchical knowledge structure. The morphology of this extracted feature is modelled using polynomial functions - this can be compared with a similar feature extracted at a different time to gain a picture of geomorphic feature development.

1. Introduction

There has been a recent and radical increase in the magnitude, speed and economics of high performance computing which has unlocked potential for computationally intensive analysis of a geographical nature. Amongst the generic applications that are set to benefit from this increased capability are artificial intelligence techniques, replacing conventional modelling tools (Openshaw & Abrahart 1996). Artificial intelligence itself has received an explosion of interest in the last five years and it is apparent that constituent areas such as expert systems will be integral in the evolution of the next generation of GIS (Fischer 1994). In Moore et *al.* (1996), a conceptual outline of an expert system was put forward for coastal zone management, an area in which there has been very little research compared with other disciplines in the geosciences. It follows that the application of expert systems to coastal zone management is unique. The expert system, COAMES (COAstal Management Expert System), strives to integrate knowledge and data into an object-oriented structure, whilst keeping the inference engine and knowledge base components of the expert system as separate entities. This provides a consistent platform to which the coastal zone manager can proffer queries and hypotheses, using the output and a holistic approach to gain a better understanding of the coast. Since this conceptual outline, the initial efforts in building COAMES have concentrated on developing a prototype covering a narrow domain in coastal expertise. This method of rapid prototyping is expedient where there is a high degree of uncertainty in the specification (Fedra & Jamieson 1996). The area of application is coastal geomorphology, more specifically the identification of beach features on a stretch of rapidly eroding coast in Eastern England (Holderness). Firstly, this paper details the preparation of digital elevation models (DEMs) of the study area through digital photogrammetric methods, before

outlining the structuring and processes that operate within COAMES. This is done with specific reference to the operation of 'intelligent' ground control points and implementation of morphometric functions held as objects within the structure. These are used to locate and delineate a specific geomorphological feature. Finally, there is a discussion on such issues as error and uncertainty, scale and modelling structures.

2. Background

2.1 Expert Systems

Expert systems can be regarded as the most mature products to emerge from the field of artificial intelligence (Raggad, 1996), dating back to the mid-1960s. A representative definition states that expert systems "....advise on or help solve real-world problems requiring an expert's interpretation and solve real-world problems using a computer model of expert human reasoning reaching the same conclusion the human expert would reach if faced with a comparable problem." (Weiss & Kulikowski, 1984). There has been much research into the use of expert systems in geography. However, progress has been slow when compared to other subject areas, mostly due to the complex nature of geospatial problems (Fischer, 1994). Having said this, the potential of expert systems is great, based on the extent to which they have been adopted in a multidisciplinary context (Durkin 1996). Indeed, very recently, expert systems have proved to be valuable in another environmental discipline, geology, where the volume of data and the complexity of processing means that 3D analysis needs computer assistance. Also the field is sufficiently huge that 'few individuals have mastery over the whole' (Ferrier & Wadge, 1997). There have been very few expert systems with a coastal application. The Ocean Expert System (Dantzler & Scheerer, 1993; Scheerer, 1993) was developed for tactical oceanography, to acquire, interpret and manage oceanographic information. A main consideration of the system was to exploit incomplete and uncertain coastal environmental information, predominantly through the Dempster-Schafer theory of belief.

2.2 Object Orientation

It has been said that there are three conceptual models to represent knowledge in an expert system - rule-based, frame-based and blackboard architecture (Kartikeyan, Majumder & Dasgupta, 1995). Historically, the rule-based model has been the most popular, though what is of interest here is the frame-based or object oriented model. Raper and Livingstone (1995) have outlined a rationale for using object-oriented techniques: it has been argued that an object-oriented paradigm (where reality is modelled through the attributes and functions relating to objects) makes considerable progress towards letting the application domain uniquely define the form of the computer model (Raper & Livingstone 1995). Conventionally in environmental modelling, the representational basis of a GIS, for example, is often allowed to drive the form and nature of the model. Ferrier & Wadge (1997) also explore avenues of possibility with object orientation, reasoning that it provides a means of structuring more complicated types of knowledge base than rule based systems.

2.3 Coastal Zone Management

The coastal zone is a unique environment where conflicting interests meet; developmental, recreational, industrial (e.g. in mineral extraction) and conservational (DoE 1995). Management is a question of reconciling these differing viewpoints. Figure 1 portrays the sociological side of coastal zone management, which enables a look at the role of an expert system in a wider context. From a sociological point of view, the coastal zone manager liaises with the coastal zone stakeholders, each having their own concerns and applications. These stakeholders will almost certainly be a fount of coastal knowledge in themselves, which they can impart to the expert system, possibly via the Internet. The conflicting applications of the stakeholders are weighed up by the coastal zone manager and fed into the system (Fig.5) via a dialogue. Based on this, the relevant knowledge and data is invoked, inferred with reference to the initial query and decision support output returned for assessment. If the output is acted upon, then cyclical monitoring of the resultant situation in the coastal zone takes place. Through use of the expert system, the manipula-

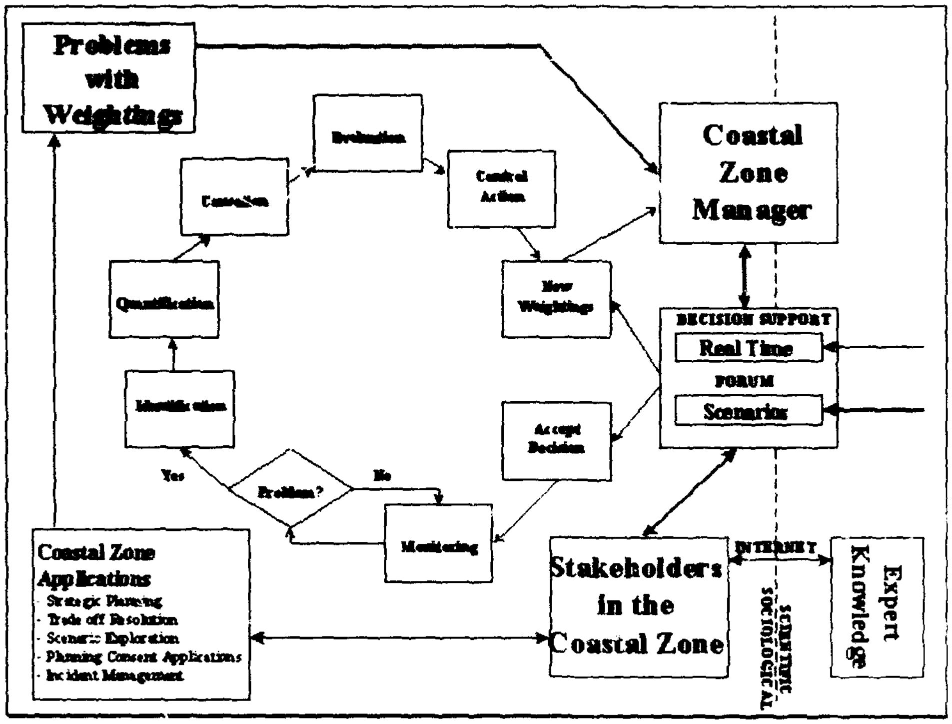


Fig.1: *The sociological component of coastal zone management*

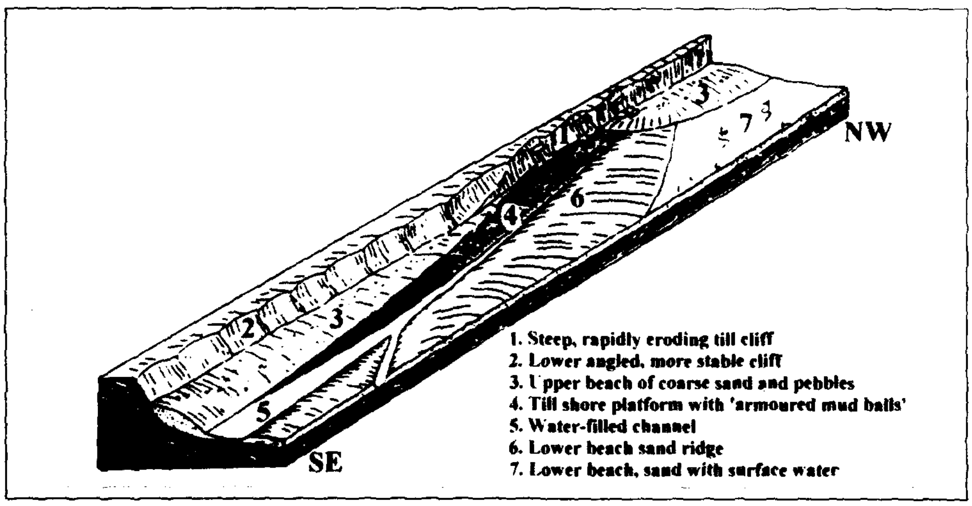


Fig.2: *The characteristic features of a Holderness ord (from Pringle 1985).*

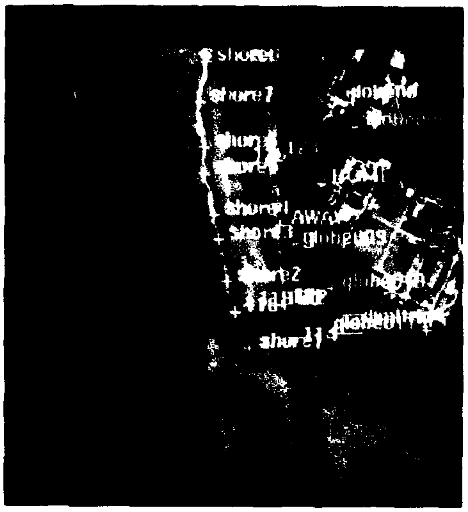


Fig.3: *Aerial photograph of the Holderness coast near Easington with superimposed ground control and tie points. Taken 26/10/96*

tion of spatial and aspatial data can be seen as a means by which effective coastal zone management can be aided. The role of such data and knowledge is to form a comprehensive platform from which informed and optimal decisions can be made on matters pertaining to the coastal zone. This is one of the main reasons why a system such as COAMES is of potential value.

2.3.1 Geomorphological Background of the Study Area

The geomorphic application chosen for this prototype reflects the conservational / natural side of coastal zone management. The area of study is the Holderness coast in northeast England, which is backed by glacial till cliffs and subject to a very rapid rate of erosion (1.2m/yr). This erosion is even more rapid where there are low sections of beach, exposing areas of till platform. These are associated with composite ridge-type beach landforms called ords, the structure of which is shown in Fig.2. These landforms migrate along the direction of longshore drift (Pringle, 1985).

Fig.4: *DEM covering part of Figure 3. The salient elements of the beach can clearly be seen, from the sand ridge on the left, through the darker till platform to the steeper upper beach banking the till cliffs*

3. Methodology

3.1 Using Digital Photogrammetry to Derive the DEM

Two aerial photographs were chosen so that the derived area of stereo overlap captured the distinct elements of one of the ords, and also so that the same area of the coastal strip was available at another time for future processing. The two times chosen were October 1996 (see Figure 3 for example) and April 1997, theoretically covering the period of most radical geomorphological change.

After prerequisite scanning, the photos were used as input into ERDAS Imagine's digital photogrammetry module, ORTHOMAX (for an overview of digital photogrammetry see Petrie 1996). Firstly, the precise positions of the two photographs in modelled computer space were pinpointed through the digitisation of their respective fiducial marks and correction for camera distortion (*inner orientation*). A further stage (relative orientation) orients the two photographs relative to each other through the identification of the same salient features (tie points) on both. The last stage of orientation is the modelling of the stereo pair to real ground co-ordinates in Latitude-Longitude or National Grid format and altitude (absolute orientation). A good spread of these co-ordinates (or ground control points) is advised across the area of stereo overlap for the optimum photogrammetric model. These known points were derived from surveyed benchmarks

and differential Global Positioning System (GPS) surveys, some of which were undertaken in conjunction with the aerial photography sorties. Associated with each ground control point measured was a description of the topology of the beach features there. It is this that is used by the expert system to locate salient elements on the beach from the DEM, which was constructed itself after stereo matching of the stereo pair. Figure 4 shows the DEM for October 1996 overlain with an orthorectified photograph (adjusted to ground co-ordinates).

3.2 Construction of the Expert System

The design of the expert system was true to the original schematic as set out in Moore *et al* (1996), which is shown in Figure 5. Briefly running through the elements and processes that underlie COAMES, an initial query prompts the interface to extract the operative words and passes them to the inference engine, which performs logical processes (e.g., induction, deduction) to select data, knowledge and models appropriately. The last role of the inference engine is to select an appropriate method for visualising the results of the query.

3.2.1 The Hierarchical Knowledge Structure

Figure 6 displays the configuration of the class structure of the geomorphological prototype COAMES. Classification involves the assignment of individual occurrences defined on the basis of selected attributes or functions. All classes will have specific attributes unique to themselves (Laurini & Thompson 1992). For instance, the raster subclasses slope, aspect and convexity are defined by their attributes and functions; these are encapsulated within the class definition. In addition, they inherit all the elements of the raster

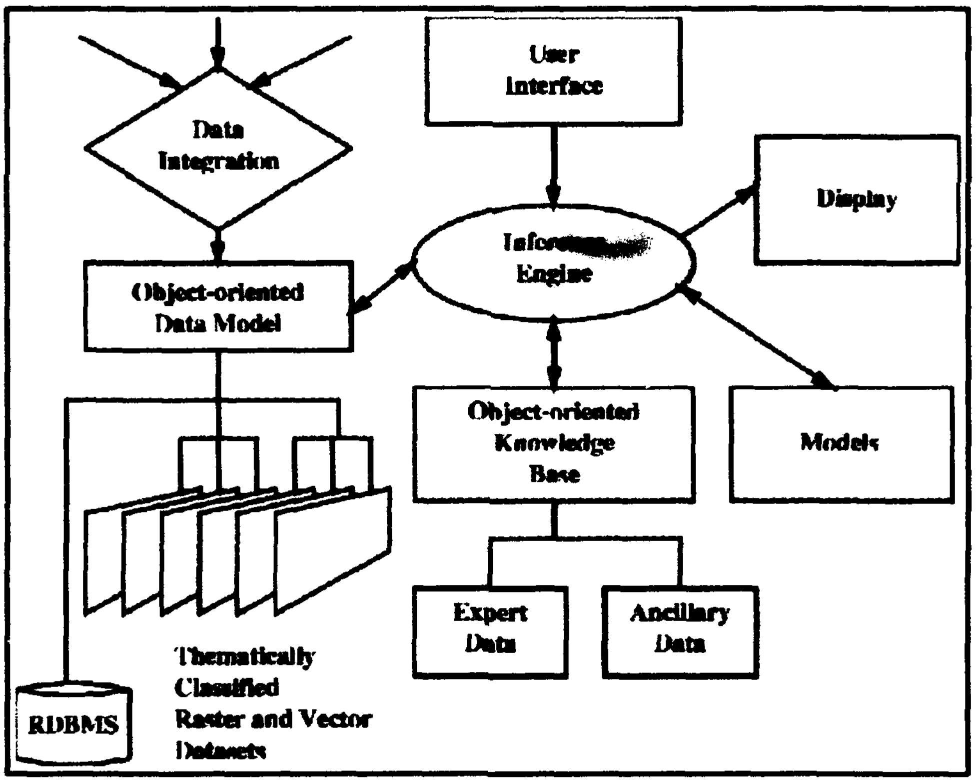


***Fig. 5:** The configuration of COAMES (from Moore et al 1996)*

superclass, such as a 2D-array data structure. In a further capability of object oriented structure, objects can be seen as communicating with other objects by passing messages that they can either accept or reject. As will be seen later, this is particularly useful for knowledge representation (Tello 1989)

3.2.2 Inference

From an initial user input (e.g. track movement of upper beach within an ord from time T[illegible] P1), a very primitive natural language process extracts words based on comparison with the contents of all the subclasses under class 'Terms' to gain coast-specific terms ('shingle', 'beach' etc), context terms ('next to', 'in' etc), temporal-specific terms ('GMT', 'low tide' etc) and landform names. Certain words (e.g. 'ord') are used to trigger or invoke a set of knowledge, in this case based on the topology between beach features held in Figure 2. The specific set of rules and facts that comprise this knowledge is itself arranged in a hierarchical object fashion. This tree is descended through a forward chaining process, initially to restrict the operation of rules to those covered by the user's query (effectively training the hierarchy for ground control point processing). For instance, the first condition asks if the user's query is concerned with cliffs. If so, then that condition is flagged 'true', which is noted by the inference engine. Subsequently, the inference engine uses information encapsulated within the original object rule to point to the appropriate object in the next tier in the hierarchy, which is to look for evidence that the cliff is steep. Also encapsulated in the rule structure is a report, which is different for each outcome as dictated by the inference engine. For instance, having ascertained that the user's query concerned cliffs, the report corresponding to 'true' would be printed out to the user: "At the base of a cliff..is it steep?" A steep cliff would indicate the edge of an ex-

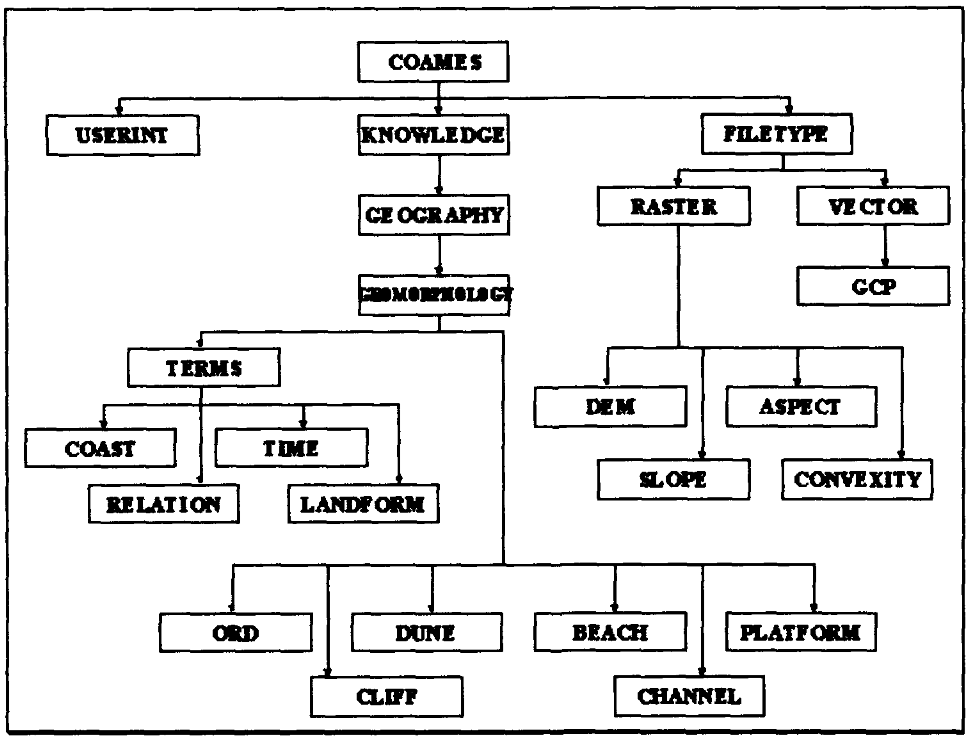


Fig.6: *The object-oriented hierarchical structure of knowledge and data in the COAMES prototype*

posed till platform, though if there is no evidence to suggest this then the inference engine would point towards a stable (and lower-angled) cliff rule. This would indicate protection from the upper beach. Again, the applicable condition is marked 'true'. After fully descending the hierarchy, the process is repeated with ground control point data (each xyz GPS point surveyed has associated topological information that further defines its position), though movement is restricted to the flagged areas. If the ground control point meets the criteria defined by the user's original query (i.e. if it in some way defines the feature to be isolated), then the associated co-ordinates are stored and used to define a region with the help of a function encapsulated in the geography class. Used in this way, the ground control points can be seen to be intelligent. Within this zoomed in region, morphometric measures (Evans, 1972; Wood, 1997) encapsulated in the geomorphology class are used to isolate the feature more specifically by using the appropriate thresholds of altitude, slope, aspect and convexity for particular land forms. These thresholds are held in the same rule structures described above within unique morphometric rule hierarchies. The result of this can be seen in Figure 7.

What must be stressed about this object-oriented expert system is that the inference method is kept separate from the knowledge base. Traditionally, the knowledge base has manifested itself as a long series of IF-THEN statements where action is taken if a certain condition is met. This exhaustive approach results in the knowledge base and inference engine being closely entwined (i.e. the action is the task of the inference engine). Ideally, the knowledge base should not be so 'hard-wired' into the system, as it may need to be modified to meet specific demands. This is best done as a separate entity (Moore *et al* 1996).

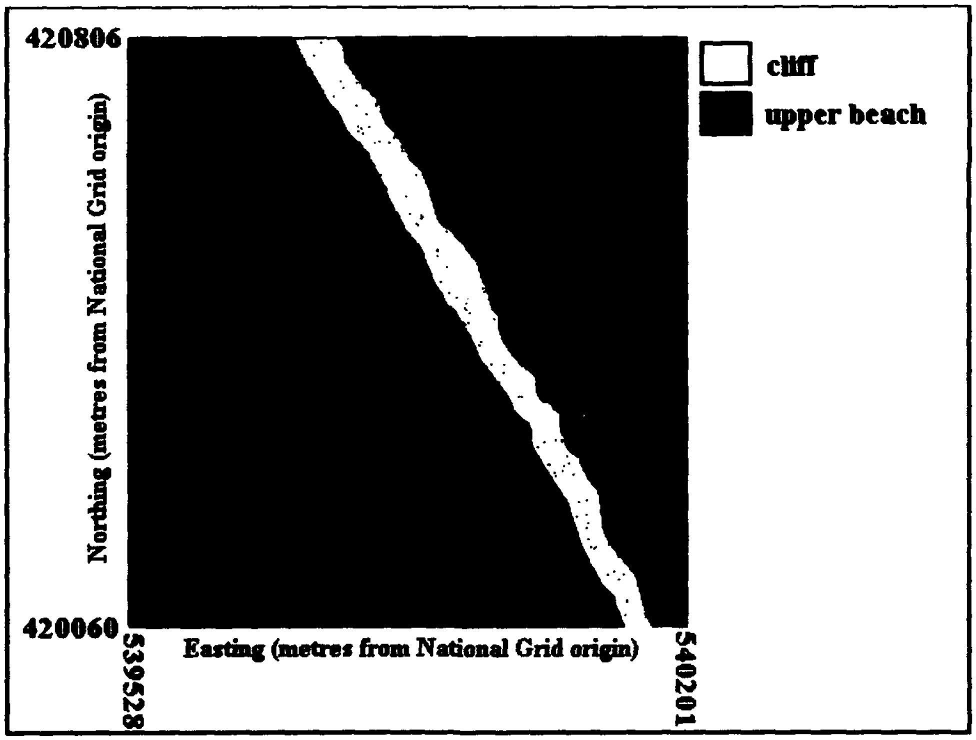


Fig.7: *The isolation and extraction of the cliff and the upper beach from the study area on the basis of intelligent ground control points and morphometric parameters driven by the COAMES expert system*

5. Discussion

5.1 Error and uncertainty

With a coastal zone management system and indeed any expert system in a commercial or academic environment, users will need to know how much confidence to attach to any output and where the confidence limits may lie. No decision made will blindly rely upon output from this expert system. Therefore, incorporation of error analysis is extremely important in COAMES' structure. Burrough (1986) identifies three broad groups of error source, which were discussed with reference to COAMES in Moore *et al* (1996). In the case of rules, for example, there is a great deal of uncertainty in defining morphometric thresholds. It would be easy enough to say that upper beaches have a slope of between 3 and 6 metres, though there are cases that fall outside this. This potential error needs to be represented in the system. It follows that some measure of the quality of results is essential for the future development of COAMES. There are a few economic and practical reasons for this (Burrough, 1992; Miller & Morrice, 1991; Moon & So, 1995). The most common error modelling methods include Bayes' Theorem (a probabilistic approach, calculating uncertainty about the likelihood of a particular event occurring, given a piece of evidence – Srinivasan & Richards, 1993; Moon & So, 1995; Skidmore et al 1996), Dempster-Schafer theory of belief functions (can be used where evidence is lacking, embodying the representation of ignorance in probability theory - Scheerer, 1993; Moon & So, 1995; Ferrier & Wadge 1997) and fuzzy logic – Zhu et al 1996, Ferrier & Wadge 1997). Fuzzy logic has been used extensively for the processing of non-crisp terms such as 'good', 'fair' and 'poor' (see Brimicombe's (1996) work with linguistic hedges). This method is potentially valuable for further development of this prototype in the processing of user queries and the quantifying of terms such as 'steep' and 'stable' cliff.

5.2 Modelling Paradigms

It is worth considering how time and space is modelled. Raper and Livingstone (1995) propose modelling within a space-time paradigm, or in relative space (4D). It is a spiral model where since return of a flux is not to the same time, it cannot be to the same place. The conventional GIS method is to have time slices based in 'absolute space'. An example of absolute space is a design where events affecting objects create 'versioned objects' so that temporally different versions of the same object can exist. (Wachowicz & Healey 1994). The relative space way of modelling can facilitate the execution of theories about relationships between 4D space-time phenomena as well as spatio-temporal interpolation.

5.3 Other Considerations

Wood (1997) divided the methods of DEM analysis into extraction and also *a priori* means. Within this group, data sources such as classified aerial photography could be used as the isolation means instead of extraction methods. Moore *et al* (1996) investigated interfacing to models from COAMES (with specific reference to nutrient exchange through a coupled pair of models). For this geomorphological prototype, the results could be input into a cliff erosion model, aiding forecasts of erosion, which itself could be linked with the important sociological element of coastal zone management (lc;s of valuable land).

5. Conclusion

This paper describes the development of a prototype version of COAMES, which represents a pioneering application of expert systems to coastal zone management. In the philosophy of COAMES, this prototype study is seen as a building block that can be added to; this is allowed by the existing formulation of the surrounding infrastructure as specified by Moore *et al* (1996) in Figure 5. Now that this initial step has been made, subsequent efforts will include the investigation of temporal and spatial change relating to a feature and linking the findings with explanatory data (e.g. wave data, suspended sediment data etc.). An incorporation of error and uncertainty is of high priority due to the need to establish the validity of the system. To be an effective coastal zone management expert system, it is important to bring in sociological and legislative knowledge. Ultimately for this prototype, though, the consideration of the ord landform as a whole is an issue, as a

means of testing the expert system's ability to prove or disprove the theory.

Acknowledgements

This research was developed within the Land Ocean Interaction Study (LOIS). The authors would like to thank Karla Youngs, Ada and George Pringle for helping with the surveying, the NERC Geophysical Equipment Pool / Ashtech Europe for loan of the GPS equipment and technical support, and the NERC Airborne Remote Sensing Facility for the acquisition of aerial photography.

References

A.J.Brimicombe, "A Universal Translator of Linguistic Hedges for the Handling of Uncertainty and Fitness-for-Use in Geographical Information Systems" in *Proceedings 4th National Conference GIS Research UK*, (University of Kent, 1996), p.143-146.

P.A.Burrough, *Principles of Geographical Information Systems for Land Resources Assessment*, (Oxford: Clarendon Press, 1986)

P.A.Burrough, "Development of Intelligent Geographical Information Systems" *International Journal of Geographical Information Systems* (1992), 6: 1: 1-11.

H.L.Dantzler & Scheerer,D.J., "An Expert System for Describing and Predicting the Coastal Ocean Environment" *Johns Hopkins APL Technical Digest* (1993), 14: 2: 181-192.

Department of the Environment (DoE), *Coastal Planning and Management: A Review of Earth Science Information Needs*, (HMSO, 1995).

J.Durkin, "Expert Systems: A View of the Field" *IEEE Expert* (1996), 11: 2: 56-63.

I.S.Evans, "General Geomorphometry, Derivatives of Altitude and Descriptive Statistics" in Chorley,R.J.(ed.) "Spatial Analysis in Geomorphology", (London: Methuen, 1972), p17-90.

K.Fedra & Jamieson,D.G., "The 'WaterWare' Decision-Support System for River-Basin Planning. 2. Planning Capability" *Journal of Hydrology* (1996), 177: 177-198.

G.Ferrier & Wadge,G., "An Integrated GIS and Knowledge-Based System as an Aid for the Geological Analysis of Sedimentary Basins" *International Journal of Geographical Information Systems* (1997), 11: 3: 281-297.

M.M.Fischer, "From Conventional to Knowledge-Based Geographic Information Systems" *Computers, Environment and Urban Systems* (1994), 18: 4: 233-242.

B.Kartikeyan, Majumder, K.L & Dasgupta,A.R., "Expert System for Land Cover Classification" *IEEE Transactions on Geoscience and Remote Sensing* (1995), 33: 1: 58-66.

R.Laurini & Thompson,D., *Fundamentals of Spatial Information Systems* (London: Academic Press, 1992).

D.R.Miller & Morrice,J., "An Expert System and GIS Approach to Predicting Changes in the Upland Vegetation of Scotland" in *GIS / LIS '91 Proceedings* (Bethesda MD: ASPRS, 1991), 1: 11-20.

W.M.Moon & So,C.S., "Information Representation and Integration of Multiple Sets of Spatial Geoscience Data" in *International Geoscience and Remote Sensing Symposium* (Piscataway NJ: IEEE, 1995), 3: 2141-2144.

A.B.Moore, Morris,K.P. & Blackwell,G.K., "COAMES - Towards a Coastal Management Expert System" in *Proceedings 1st International Conference on GeoComputation* (University of Leeds, 1996), 2: 629-646.

S.Openshaw & Abrahart, R.J., "Geocomputation" in *Proceedings 1st International Conference on GeoComputation* (University of Leeds, 1996), 2: 665-666.

G.Petrie, "Developments in Digital Photogrammetric Systems for Topographic Mapping and Geoscience Applications", Paper given at *Remote Topographic Mapping for Geoscience*, (British Geological Survey, Keyworth, 1996).

A.W.Pringle, "Holderness Coast Erosion and the Significance of Ords" *Earth Surface Processes and Landforms* (1985), 10: 107-124.

B.G.Raggad, "Expert System Quality Control" *Information*

Processing and Management (1996), 32: 2: 171-183.

J.Raper & Livingstone, D., "Development of a Geomorphological Spatial Model using Object-Oriented Design" *International Journal of Geographical Information Systems* (1995), 9: 4: 359-383.

D.J.Scheerer, "Reasoning under Uncertainty for a Coastal Ocean Expert System" *Johns Hopkins APL Technical Digest* (1993), 14: 3: 267-280.

A.K.Skidmore, Watford,F., Luckananurug,P. & Ryan,P.J., "An Operational GIS Expert System for Mapping Forest Soils" *Photogrammetric Engineering and Remote Sensing* (1996), 62: 5: 501-511.

A.Srinivasan & Richards,J.A., "Analysis of Spatial Data using Knowledge-Based Methods" *International Journal of Geographical Information Systems* (1993), 7: 6: 479-500.

E.R.Tello, *Object-Oriented Programming for Artificial Intelligence*, (Addison-Wesley, 1989).

M.Wachowicz & Healey, R.G., "Towards Temporality in GIS" in Worboys,M. (ed.) *Innovations in GIS 1* (London: Taylor & Francis, 1994) pp.105-115.

S.M.Weiss & Kulikowski,C.A., *A Practical Guide to Designing Expert Systems* (Totawa NJ: Rowman and Allenheld, 1984).

J.Wood, "Some Implications for GIS of a Multi-Scale Representation of Surfaces" in *Proceedings 5th National Conference GIS Research UK*, (University of Leeds, 1997), p.9-13.

A-X.Zhu, Band,L.E., Dutton,B. & Nimlos,T.J., "Automated Soil Inference Under Fuzzy-Logic" *Ecological Modelling* (1996), 90: 2: 123-145.

Eliciting Knowledge with Visualization - instant gratification for the expert image classifier who wants to show Rather than tell

Paul Crowther and Jacky Hartnett
Artificial Intelligence and Spatial Systems Research Group
Department of Computing
University of Tasmania
P.Crowther@utas.edu.au

Presented at the second annual conference of GeoComputation '97 & SIRC '97, University of Otago, New Zealand, 26-29 August 1997

ABSTRACT

GIS systems the world over are awash with data that experts can classify visually. This process is time consuming and costly. Expert Systems have been built which attempt to at least pre-classify images and hence speed up the process. To build these systems it is necessary to elicit information from the human expert classifiers in order to assist the classification of these many hundreds of images. Traditionally this knowledge has been captured through interview and protocol analysis. However, this required either the expert classifier to describe verbally what they were seeing or the expert systems developer (knowledge engineer) to interpret what they were being shown.

To overcome this problem, a visual knowledge acquisition tool, KAGES (Knowledge Acquisition for Geographic Expert Systems), was developed. Impetus to the development of this tool was given to our group by the need to classify many remotely sensed images of Antarctica in order to provide information on global climate change and Southern Ocean currents.

1 INTRODUCTION

This paper describes a tool for acquiring knowledge from expert image interpreters by allowing them to demonstrate their expertise. To do this the tool must work quickly in order to provide the user with rapid feedback on the knowledge acquired. KAGES was developed on a work station using Research Systems Incis (1994) IDL image processing package to overcome this problem. The matrix handling features of IDL have allowed a fast and efficient system to be developed which allows a human classifier to identify features of interest by pointing or drawing on an image displayed on a computer screen. KAGES captures the knowledge underlying the identification of these features in the form of production rules. These can include rules that describe the spatial relationships between two of the identified feature types as well as rules that identify features in terms of their spatial relationships within a group of features.

KAGES is a toolkit which provides a series of knowledge acquisition techniques including an interview manager, several graphical acquisition tools and a rule editor. The captured classifier knowledge is held in a series of knowledge bases which are then consolidated and checked for consistency and redundancy. The result is a knowledge base which can be viewed and reviewed by the human classifier.

Without exploiting the speed of current workstations, this type of computing would not be feasible since interactive graphical knowledge acquisition of visual knowledge is computationally expensive. KAGES operates on all bands of a satellite image and overlays the results on a composite image. The data structure which is being manipulated in memory is therefore an array of 1000 X 800 X 6 in the case of NOAA images. This can be even larger when more

than one version of the composite image is required, adding extra dimensions to the array structure. Since the operation is pixel-based and the results are held as a raster format, there is a large memory requirement for individual objects as well. Users require instant feedback about the results of the operation of the various tools to allow for quick verification. As a consequence, zoom and scaling functions require fast processor speed as these are required to operate on all dimensions and objects. Processor speed is also at a premium when the spatial analysis tool is operating, as all spatial relationships, including overlap, proximity and orientation, are captured.

Currently KAGES operates as a serial processor, but many of its functions are well suited to parallel processing. This is particularly true of the spatial relationship and the knowledge base consolidation tools which have a number of distinct independent functions.

2. VISUAL KNOWLEDGE

Knowledge is understanding, awareness, or familiarity acquired through education, or experience, anything that has been learned, perceived, discovered, inferred or understood and the ability to use information (Nagao, 1988).

Maps present knowledge naturally occurring in three dimensions in a two-dimensional graphic form. However, maps are produced from information from ground (and sometimes underground) survey, from images produced by sensors on aircraft or satellites, and from photographic images. Each of these could be regarded as another dimension. The problem, then, becomes one of representing n dimensional knowledge in a two dimensional form. The information is then used to produce a map showing some specific characteristics of an area (land use, soil type, geology, vegetation cover for example). To produce maps, experts use some or all of the information sources (dimensions) listed above. By using expert system approaches it may be possible to make more use of all the dimensions of information available and the interpretation of that into the knowledge of multiple domain experts.

Acquiring knowledge from multiple expert classifiers also introduces another problem, that of assigning definitions to features (Kweon and Kanade, 1994). In geography, most terms are described in natural language, but the definitions are often incomplete or open to interpretation. This interpretation may also be culturally based (Clementini et al, 1993). However, the visual definition of the feature in a graphical form is more concrete and less subject to interpretation. What it looks like defines it, rather than what it is called.

Spatial knowledge here will be defined as knowledge of entities in two dimensions (as in a map) as distinct from three dimensions where research is more directed towards vision and recognition.

McKeown et al (1989) identifies five types of knowledge used to identify specific spatial relationships. The five types are:

- Type 1 Knowledge: identifies scene primitives where a primitive is a readily identifiable object such as a road or a building.
- Type 2 Knowledge: is the knowledge of the spatial relationships between the scene primitives, for example buildings are next to roads, icebergs are surrounded by water.
- Type 3 Knowledge: defines collections of objects which form spatial decompositions within the task domain.
- Type 4 Knowledge: consists of how to combine information from type 3 knowledge.
- Type 5 Knowledge: is used to resolve and evaluate conflicting information.

This classification has become the basis of the tools and techniques described in this paper. Further sub classification has been necessary due to the characteristics of the object(s) being investigated. Hence, lines have different characteristics from areal scene primitives at the Type 1 level.

3. A VISUAL KNOWLEDGE ACQUISITION SYSTEM

KAGES consists of tools to capture knowledge of the first three types using visual tools. This overcomes some of the problems associated with verbal descriptions and

definitions. Type 5 knowledge is addressed by a module which combines knowledge gained from different image interpreters, different images and different sessions into a consolidated knowledge base. The system interacts with the user in order to resolve inconsistencies. This tool will not be discussed in detail in this paper.

An interview manager based on repertory grids and personal construct theory (Kelly, 1955) is also provided (Crowther and Hartnett, 1996). This tool was developed to test usersí reaction to both a visual and a text based tool (although it uses visual cues) and provide an alternative knowledge acquisition technique to deal with non visual knowledge. This was the tool for experts who would rather ëtellí than ëshowí.

3.1 TYPE 1 TOOL

Domain primitives are features to which an image interpreter can point and give a name. They may be point, line or areal features. Each requires different processing as each has different properties. The most simple features are point features which are subpixel in nature and which are not generally identifiable by pixel threshold signatures. These features generally are a fixed point such as a building. The expert interpreter identifies these objects by selecting an image band and pointing to their position. The name of the point feature, its location and the identification of the image on which it was defined are stored.

Areal features are two dimensional objects which may be permanent features (such as a lake) or transient (for example the contents of a field or the extent of sea ice). The user first selects the band or composite bands they use for identifying a feature. The feature is then described by an expert pointing at it and setting pixel threshold values. KAGES allows this by the use of slider bars. All pixels within the thresholds which are contiguous are then grouped (Fig 2). If the grouping agrees with the expertís idea of the extent of the feature, it is named and information about thresholds and its minimum bounding rectangle (MBR) (Chang and Jungert, 1996) is stored.

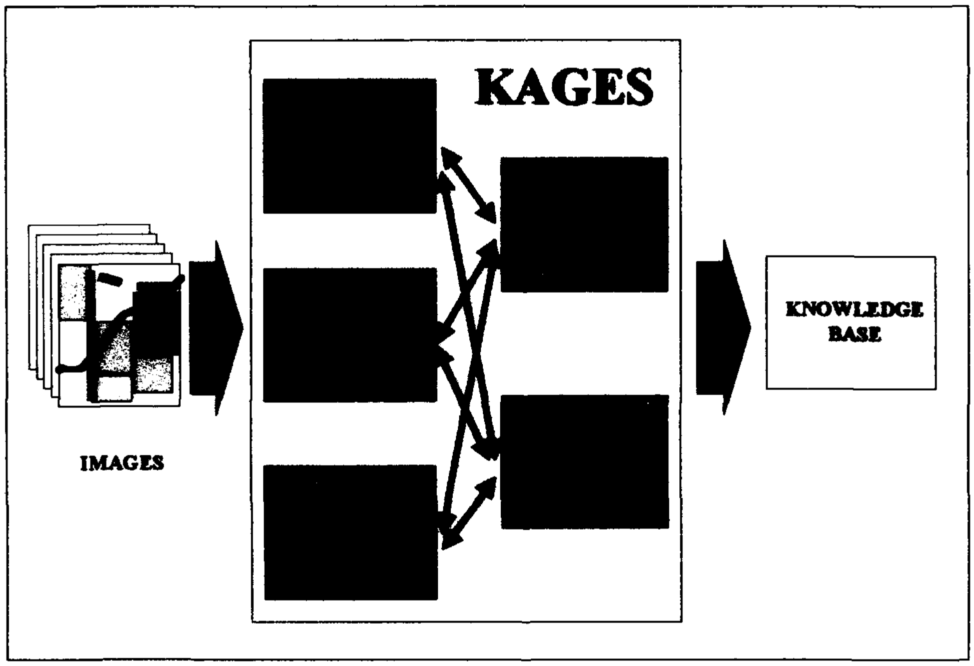


Figure 1: The KAGES system

The expert user can identify other occurrences of the object type on the training image (although it could be on other bands or band combinations). At any stage the user can ask for production rules defining object types to be generated.

Lines have proved to be the most difficult objects to deal with (Crowther and Hartnett, 1997). These one dimensional objects can be identified by either a line following algorithm or by line tracing. The later technique is necessary for lines such as municipal boundaries and other cadastral data. Information about lines is stored as either a set of raster points or as a vector, depending on how the line was acquired.

A user can choose as many examples of an object as they wish on whichever band or combination of bands they like. Once this has been done, the module which consolidate rules can be called. This filters the knowledge base combining rules with the same antecedents. The resultant rules can be fired individually with the results overlaid on band 1 (even if the objects were identified on other bands), or the entire Type 1 knowledge base can be applied.

3.2 TYPE 2 TOOL

The Type 2 tool is the spatial relationship tool which allows a user to determine the relationship between two objects (Fig 3). The two objects are shown named on the default band of an image set (usually band 1) with their MBRIs. KAGES then determines the relationships between the two objects. Objects fall into the three scene primitive types and procedures have been developed to deal with relationships between them. These relationships fall into the following categories:

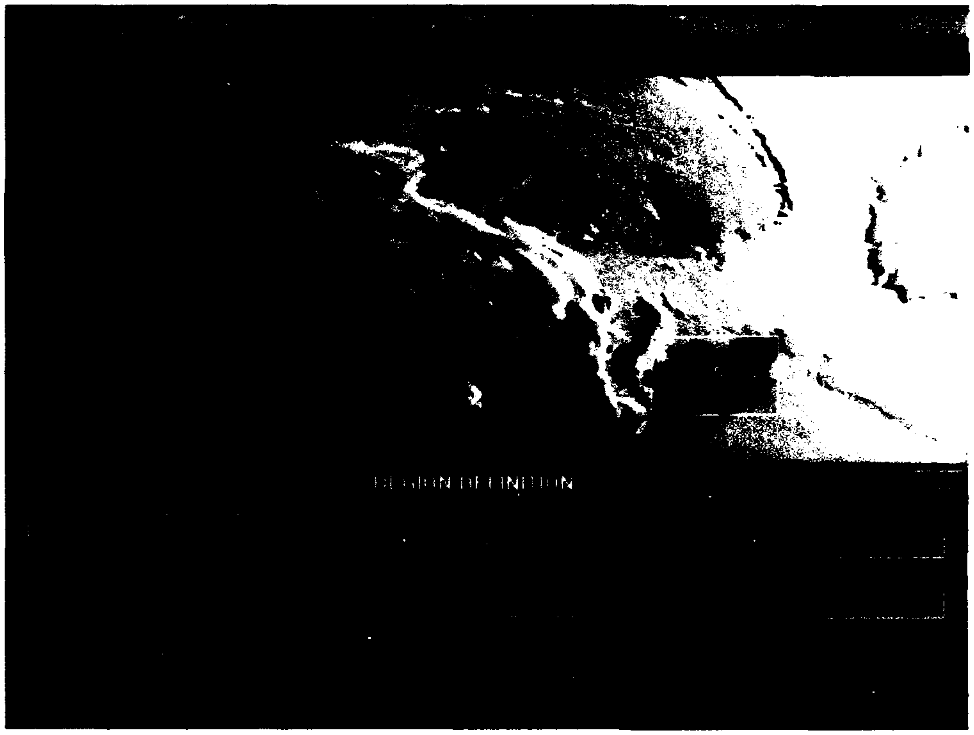


Figure 2. The per pixel Type 1 tool being used on an areal object. The object being defined in this Band 1 NOAA image of Vincennes Bay is open water. The objects Minimum Bounding Rectangle (MBR) is shown with the centroid of the MBR being marked by an O. Information on threshold values and the user entered feature name are shown in the dialogue window.

point : point	point : line	point : area
line : line	line : area	area : area

The type of relationships fall into three main categories based on Allen intervals as used by Egenhofer (1991). In all cases the following relations are calculated:

Proximity
Degree of overlap
Orientation are calculated.

The spatial relationship module first determines what the two types of objects are it is dealing with. The system then determines all possible relationships between the objects. For example in the case of two lines being selected:

Determine if the lines touch
If they touch
    Determine type of intersection
Otherwise
    Determine directional relationship
    Determine degree of parallelism

Operationally, the user is required to select two objects, which are related, from a menu and then check the results of the relationships generated by the system. This is done by the use of a simple rule editor which allows a user to remove clauses which are chance relationships and not deterministic of the relationship between the two objects.

As a further example, if a point object is compared with an areal object, the system determines the distance of the

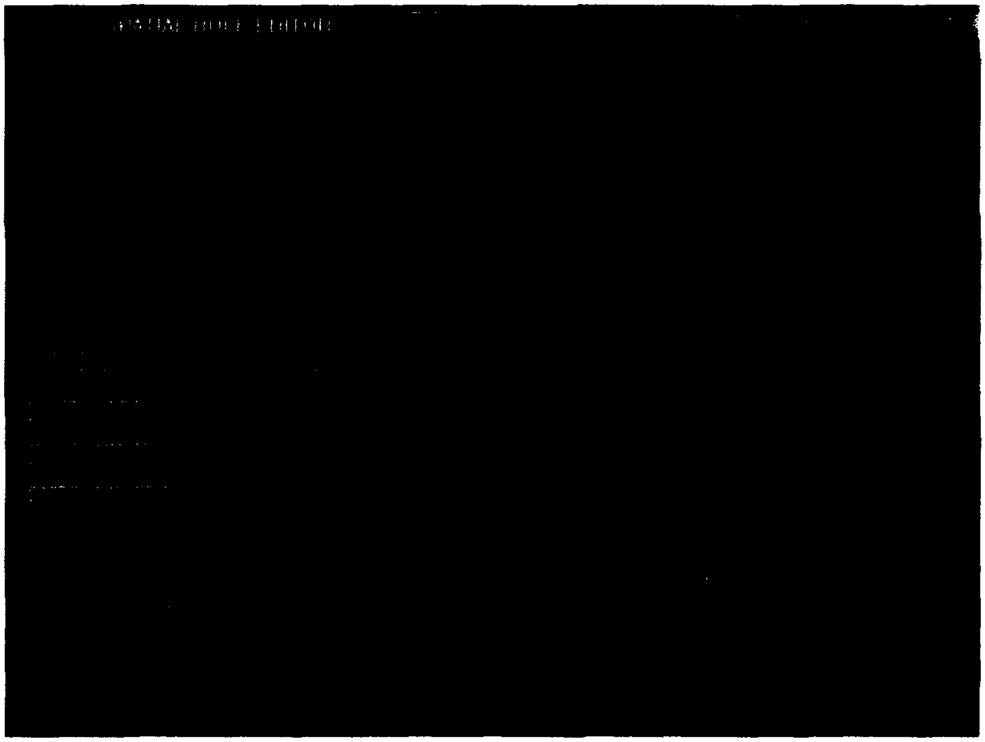


Figure 3: The spatial Type 2 tool being used to determine the relationships between a line object (a road) and an areal object (field) in a Landsat image of an area near Perth in Northern Tasmania. These are both functional primitives identified by the Type 1 tool. The relationships are shown on the image and as a rule in the rule editor window.

point from the centroid of the area, the direction of the point from the centroid (equivalent to orientation) and the relationship of the point to the boundary of the area. Possibilities are:

- Outside area and disjoint
- Outside area and touches boundary
- On the Boundary
- Inside area and touches boundary
- Inside area

3.3 TYPE 3 TOOL

The Type 3 tool allows a user to group objects defined by the Type 1 tool. Often these type of features have no natural boundary and are traced by an image identifier. This tool allows a user to trace a boundary directly on the image band they have chosen (Fig 4).

All objects which fall in the region of interest, whether fully or partially, are shown. Hence a river flowing through the region would be displayed, even though its points of termination may fall outside the region. Objects defined using other bands are also shown.

The results of the objects identified by the tool are displayed in a window labeled ìCheck Membershipî which can be manipulated by the user to remove objects which are not distinctive of the region. Once the user has completed this task a naming window is displayed. This name, the members of the region and the band used together with image identification are then stored.

4. TOWARDS A TYPE 4 TOOL

The aim of KAGES is to use an image as training data to create a rule base which can then be applied to other im-

Figure 4: The Type 3 Region Of Interest tool being used to pick up point, area and line objects using band 1 of a Landsat image. The limits of the Region Of Interests MBR are marked with a (. Paddock1 and Paddock2 MBRs are drawn with their centroids named. Points (Robyn and the cow) are named with the location of the point at the left of the name. Line objects, in this case the river, are named near their centre point.

ages. It was not intended to be an image classifier. As part of the development, a verification tool which applies rules has been built and this will do classification. This tool operates in three stages:

Apply Type 1 rules to the image

Segment the image and label individual objects

Apply spatial (Type 2) rules to the segmented image.

The result is a classified image which shows individual classified objects, areas for which no rule has fired (and hence are unclassified) and areas where conflicting rules have fired. These last two highlight image features which require further investigation by the KAGES tool.

A side effect of this development has been to allow this feature to be used as an image analyzer. As once the expert user is satisfied with the performance of the generated rules on the training image, those rules can then be applied to other images. Development of this tool turned KAGES from being just a knowledge acquisition tool into a tool with image analysis capabilities which will provide knowledge of complete scenes (Type 4 knowledge).

5. FULLY AUTOMATED OR HUMAN ASSISTED

Experience with Icemapper (Williams et al, 1997), which was developed using rules acquired by traditional interview techniques lead to the development of a system which provided a first best guess and which could then be adjusted by the image interpreter. Generally after three iterations a properly classified image was produced. This cut down the time of development of a classified image from about an hour (fully manual) to around ten minutes. This human directed system was preferred to a fully automated version based on neural networks (Kilpatrick and Williams, 1995).

KAGES was developed with this human assisted ethos in mind. The amount of time taken for a human to tell (in the form of interviews) and the knowledge engineer to interpret (into rules) took several months for Icemapper. This was a classic example of the "knowledge acquisition bottleneck" (Gaines, 1988). To speed the process up, the tool that allowed an expert to show rather than tell their expertise was developed.

The tool also gets around the problem of feedback from the knowledge engineer in that by being able to apply the rules which they have developed, the expert user can see within four seconds, the results of applying those rules. If necessary, changes can then be made and the rules reapplied.

6. KAGES AND THE NEED FOR SPEED

Speed is essential in two areas. First there needs to be sufficient processing power to handle the size and number of arrays containing the images. These are generally multi-dimensional with each band taking up a two dimensional array and the complete image set and any result images being other dimensions. For example a NOAA image of 1000 by 800 pixels with 5 bands and an array of a similar size for manipulating and displaying the results requires a minimum of 4.8 megabytes of memory. This is an underestimation of the actual memory requirement as composite bands will add further dimensions to the data structure as will other arrays which are used to store individual features and intermediate results of processing. A user requirement to compare the training image with other images will double this. The key factor is that the processing must take place sufficiently quickly for the user to respond to the feedback iteratively. Fortunately memory is cheap and the system has been run successfully on a 16 Mb machine.

The description of spatial relationships is one area where the implementation of parallel algorithms has potential to further speed up the system. We have identified the algorithms used in the current serial version of the system to have Multiple Input Multiple Processing (MIMP) (Ding and Densham, 1996) characteristics. Futhermore, to apply spatial rules for verification, all occurrences of one class of objects are compared with all occurrences of a second. In this case Multiple Input Single Process (MISP) could be used as the same procedure is used on multiple occurrences of an object type. This analysis provides exciting possibilities for future implementation.

For now, IDL routines, which allow manipulation of arrays, speed up data processing. A speed increase of about sixty (10 minutes to 10 seconds) was achieved when spatial relationship routines were converted from traditional loop processing.

7. APPLICATIONS

The system is currently being used to generate rules for comparison with those already manually determined and placed in the Icemapper sea ice identification system. The rules are comparable but are much quicker and easier to both generate and modify. Of interest, several of the rules in Icemapper can only be reproduced using the repertory grid interview tool as they are not visual in nature.

A second system currently under investigation and development involves crop identification in the North West of Tasmania. In this system the rules generated by KAGES are applied using the module currently used for verification. In this case the images are supplied by a GIS and the results will be ported back into the GIS. The results, which are so far incomplete, are being compared with the results gained from clustering techniques.

8. CONCLUSIONS

To capture visual knowledge for the use in an expert system coupled to a geographic information system, a primarily non text graphical tool is necessary. Such a visual knowledge acquisition tool has to have the following characteristics:

- It must have a graphical user interface
- It must be intuitive for the user to operate.
- It must acquire the useris knowledge by directly capturing their actions
- It must operate in real time and give instant feedback
- The rules generated must be visually verifiable

KAGES is designed to meet the user requirements but relies on powerful hardware to function successfully. The system is designed to be cross platform and work on both personal computers and workstations. Current generation personal computers with at least 16 Mb of RAM give satisfactory performance. On workstations the system has given good results and has a high level of user satisfaction.

9. REFERENCES

Chang, S. and Jungert, E. (1996), Symbolic Projection for Image Information Retrieval and Spatial Reasoning, Academic Press, London, Ch. 2.

Clementini, E Di Felice, P. and Van Oosterom, P.(1993), A Small Set of Formal Topological relationships Suitable for End-User Interaction, SSD '93 Advances in Spatial Databases, Abel, D. & Ooi, B.C. (eds), Singapore, Springer-Verlag, pp 277 - 295.

Crowther, P. and Hartnett, J. (1996), ëUsing Repertory Grids for Knowledge Acquisition for Spatial Expert Systemsí, Proceedings of the 4th Australian and New Zealand Conference on Intelligent Information Systems. Adelaide, pp 14 - 17.

Crowther, P. and Hartnett, J. (1997), ëThe Problem with Lines - Spatial Reasoning with One Dimensional Objects in a Two Dimensional Representation of Three Dimensional Spaceí, Proceedings of the 8th Joint European Conference and Exhibition on Geographic Information Systems. Vienna, pp 167 - 176.

Ding, Y. and Densham. P. J. (1996), ëSpatial Strategies for Parallel Spatial Modellingí, International Journal of Geographic Information Systems, Vol. 10, No. 6, pp 669 - 698.

Egenhofer, M. J. (1991), ëReasoning About Binary Topological Relationsí, in Advances in Spatial Databases, 2nd Symposium SSD ë91, Gunther, O and Schek, H.J. (eds), Springer Verlag, pp. 143 - 159.

Gaines, B. R. (1988), : An overview of Knowledge Acquisition and Transfer. In: Knowledge-based Systems Vol. 1., Gaines, B. R. and Boose, J. H.(eds)., Academic Press, pp. 3 ñ 21

Kelly, G.A. (1955), :The Psychology of Personal Constructs, New York: Norton.

Kilpatrick, D. and Williams, R. (1995),: Unsupervised Classification of Antarctic Satellite Imagery using Kohonenís Self-Organising Feature Map Proceedings of IEEE International Conference on Neural Networks, Perth, western Australia, Vol. 1, pp. 32-36

Kweon, I. S. and Kanade, T. (1994), ìExtracting Topographic Terrain Features From Elevation Mapsí, CVGIP:Image Understanding, Vol 59 No 2 pp 171-182.

McKeown, D. M., Harvey, W. A. and Wixson, L. E. (1989), ëAutomating Knowledge Acquisition for Aerial Image Interpretationí, Computer Vision & Graphics, Vol. 46, pp. 37 - 81.

Nagao, M. (1990), Knowledge and Inference, Academic Press, San Diego Ch. 1.

Research Systems, Inc. (1994), Interactive Data Language Version 3.6.

10. ACKNOWLEDGMENTS

The expert system development work reported in this paper has been supported by an Antarctic Science Advisory Committee Research Grant.

EpiMAN-TB, a decision support system using spatial information for the management of tuberculosis in cattle and deer in New Zealand

J.S. McKenzie[1], R.S. Morris[1], C.J. Tutty[2], D.U. Pfeiffer[1]
Dept of Veterinary Clinical Sciences,
Massey University,
Palmerston North, New Zealand.

[2] 76 Cook Street,
Palmerston North, New Zealand

Presented at the second annual conference of GeoComputation '97 & SIRC '97, University of Otago, New Zealand, 26-29 August 1997

Abstract

EpiMAN-TB is a decision support system that is being developed as a tool to assist the development and evaluation of TB control strategies, and the management of the TB control program in New Zealand. It takes the form of an epidemiological workbench of tools to support TB control decisions made by field veterinarians and farmers. These tools include: the prediction of possum TB hot spots, classification of farms according to TB risk, evaluation of possum TB control strategies at the level of individual farms and at regional level. EpiMAN-TB comprises a database, map display tools, simulation models of the spread of TB between possums at farm and at regional levels, and decision aids based on expert systems. It utilises spatial information relating to vegetation cover, topography and farm boundaries, plus TB history and management information for individual farms.

1. Introduction

Tuberculosis (TB) in cattle and deer is a problem of national concern to the New Zealand pastoral industries due principally to its negative impact on export markets. A national TB control program has been in place since the 1970s. However, efforts to control the disease in farmed animals are hampered by the presence of TB in the brushtail possum (*Trichosurus vulpecula*). Infected possum populations are the major source of TB infection for cattle and deer in New Zealand. Thus control of the disease in farmed anin[illegible]lling the disease in infected possum populations.

TB control strategies used to date have successfully reduced the total number of infected possums and consequently the total number of infected cattle and deer. However, they have not been successful in eradicating TB from possum populations, hence continued control of these populations is required to maintain the incidence of TB in farmed animals at a low level. We believe that more effective control strategies can be developed through an integrated approach involving the use of scientific information and field data on the epidemiology and spatial distribution of the disease in possums and the use of models to compare the effects of different strategies.

The spatial distribution of TB in possums is clustered at three scales. It is present in possum populations only in certain regions of New Zealand, and within these affected regions certain farms or groups of farms have a TB problem while others have no (or very infrequent) infection. The smallest unit of clustering is possum denning areas occupying as little as 0.25-0.5 hectares (Hickling, 1995). Field research indicates that TB clusters in possums fall into two categories: endemic and sporadic. At the site of endemic clusters, TB spreads well between possums and remains in the same location for many years. These are colloquially known as 'hot spots'. Sporadic clusters remain

for only short periods of time as TB does not spread so well between possums and has not become firmly established at the site. It is highly likely that TB is maintained in possum populations at the sites of endemic clusters even after the population density has been reduced to a low level by control measures. It is the perpetuation of TB at such locations that means possum populations must continually be kept at a low level to reduce the risk of TB spreading from possums to farmed cattle and deer. This is a major expense to the industry, and the ability to target control measures at areas where the effect is likely to be greatest would improve the efficiency with which possum control resources are used.

Research has identified features of habitat that can be used to predict the potential location of endemic, sporadic and negative TB sites. Endemic clusters are more likely to occur on flat or gently sloping land with large-diameter trees that provide well-enclosed den sites. Sporadic clusters are more likely to occur on flat or gently sloping land with taller trees, but which do not have multiple enclosed den sites available. Negative TB sites are more likely to occur on steeper slopes covered in scrub. This information can be used at the scale of an individual farm to predict high risk areas within the farm, or at the scale of a region to predict farms within the region that have vegetation patterns which increase the likelihood of possums on the farm being infected with TB. Such information can then be used to assist with the formulation of TB management strategies at either farm or regional level. This paper describes a decision support system, EpiMAN-TB, which is being developed as a tool to use this spatial information in the development and evaluation of TB control strategies and in the management of the TB control program in New Zealand.

EpiMAN-TB

2.1 Description

EpiMAN-TB is a decision support system designed for the use of TB managers, mostly at the field level. It will assist in the formulation of TB control programs for farms and larger areas, and will allow evaluation of alternative control programs. It will make possible comprehensive forms of assessment of progress in TB control at district, regional or national level, and it will permit policy assessments to be made for potential new control methods. This decision support system provides a way of integrating the current state of knowledge on TB and possums into disease management decisions, with the assumption that better decisions will be made. It also provides a way of incorporating sophisticated information processing technology into the day-to-day decision making process in a form that is simple to use.

It is a stand alone system that will be used on PCs by TB management field staff throughout the country. Emphasis has been placed on it being a generic tool that can be put into any office, and a deliberate effort has been made to not be dependent on any particular commercial GIS or database management software. GIS functionality and other essential features are provided within a range of standard SQL compliant database programs. The generic nature of the software also allows it to be adapted to manage other endemic diseases that have a strong spatial component in their epidemiology.

2.2 Structure

EpiMAN-TB comprises a database, map display tools, a simulation model of TB in possums, and decision aids based on expert systems, as illustrated in figure 1. Database information required to run EpiMAN-TB relates to farm ownership, animal numbers and TB status of cattle and deer on the farms. This information is currently available in databases which are either owned or managed by MAF Quality Management (MQM). Farm information is obtained from Agribase which is a national database of farms in which each farm is uniquely identified. Agribase contains basic property ownership and land use information plus locational information that facilitates the production of farm maps. TB status information is obtained from the National Livestock Database (NLDB). This database contains a history of TB testing results for most farms in New Zealand. Farms are identified by the Agribase farm identification

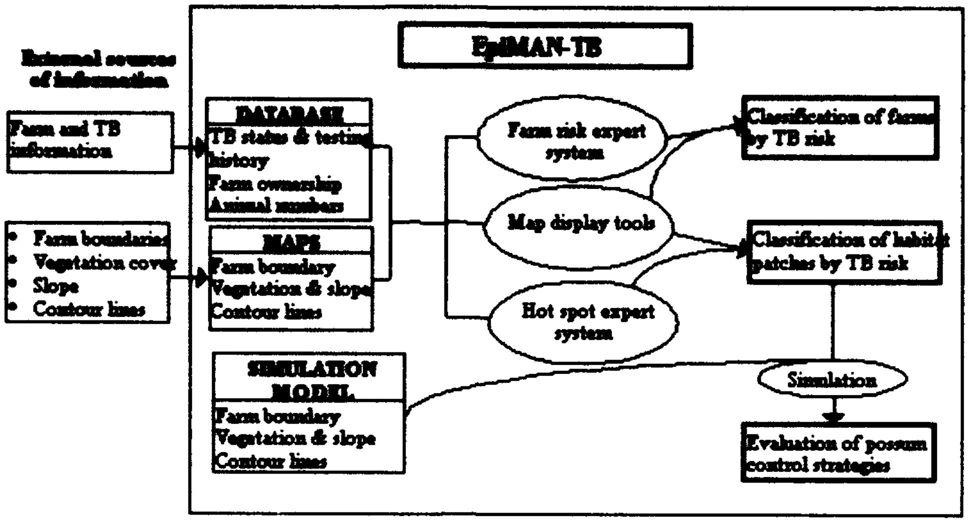


Figure 1. An overall view of the structure of EpiMAN-TB

number so that information in the two databases can be linked.

At present, the database information is extracted from the original two databases and is maintained separately within EpiMAN-TB. As MQM staff are likely to be the major users of this system, the establishment of a live link between EpiMAN-TB and the original databases will be explored. This will provide for more efficient use of computer space, and ensure that the information in EpiMAN-TB is as current as that in the original databases.

EpiMAN-TB does not include sophisticated spatial manipulation tools. Users require access to map data that has already been processed with a GIS into the form required by EpiMAN-TB. Such processes include creation of digital elevation models, overlay analyses, map generalisation, and others. The geographic tools programmed into EpiMAN-TB are predominantly to display map information in different ways customised to the users' needs, and to undertake various analytical procedures. The map information required by EpiMAN-TB is: farm boundaries, rivers, roads, vegetation cover, slope of the land and contour lines. Details of information in each of these maps follows.

Farm boundary maps are vector maps outlining the boundaries of individual farms, each with an associated farm identification number from Agribase. Vegetation and slope maps are both raster images with 40 meter pixels. Vegetation classes were derived from a SPOT multi-spectral satellite image. The classes of vegetation are: podocarp-broadleaved forest, beech forest, pine forest, scrub, willows, shelter belts, swamps and pasture. A SPOT multi-spectral image was chosen as this provided an appropriate spatial resolution of 20 meters with adequate, though somewhat limited, spectral resolution. SPOT MS imagery is the best currently available in New Zealand with good spatial resolution. As information with higher spatial and spectral resolution becomes available in the future, enabling more detailed vegetation maps to be produced, these can be incorporated into EpiMAN-TB, if it is found that the greater differentiation of vegetation improves the accuracy with which possum TB hot spots can be predicted.

2.3 Functions

Users are able to select from a number of different tasks available within the software, depending upon their specific need. Tasks are outlined in figure 2 and each task is described in more detail below.

2.3.1 Hot spot prediction

Having the ability to predict hot spots, or the location of habitats where TB is likely to be endemic in possums on a farm, helps develop a TB management plan for the farm. These high risk areas can be targeted for more intensive possum control efforts, and/or can be avoided in a cattle or deer grazing program.

Prediction of possum TB hot spots utilises farm boundary, vegetation and slope information. This task can be run for an individual farm or for a small area including a number of farms. Farms are identified by entering the farm identification number which brings up the farm plus a buffer of 100 meters around the boundary. An alternative area can be selected interactively by the user. This defines the geographic boundaries for the vegetation and slope map which is used in the prediciton process. Vegetation cover is represented in 40 meter pixels and the hot spot expert system is then run for all cells in the selected area, assigning one of three TB risk categories to each cell. Risk categories are high, medium and low. EpiMAN-TB outputs a map shading each cell according to its risk category. Contour lines are drawn on the map to provide some contextual information to help users identify landmarks. Hard copy of this map can be given to a farmer to take away and use to develop a TB management program.

2.3.2 Possum control strategy evaluation

i) Farm or small area control

Having identified areas of habitat where the risk of a TB hot spot is high, alternative possum control strategies can be compared for their influence on reducing the prevalence of TB in the possum population. This can be done in EpiMAN-TB by means of a simulation model of TB in possums, PossPOP, which can be run for a single farm or a small group of contiguous farms. PossPOP is a geographic model representing the ecology and infection dynamics of wild possum populations, which includes natural stochastic variation, spatial (spatial heterogeneity and autocorrelation) and temporal (seasonal and cyclical effects) effects (Pfeiffer et al, 1994). PossPOP uses a real vegetation map for the area of interest in the simulation to better represent real-

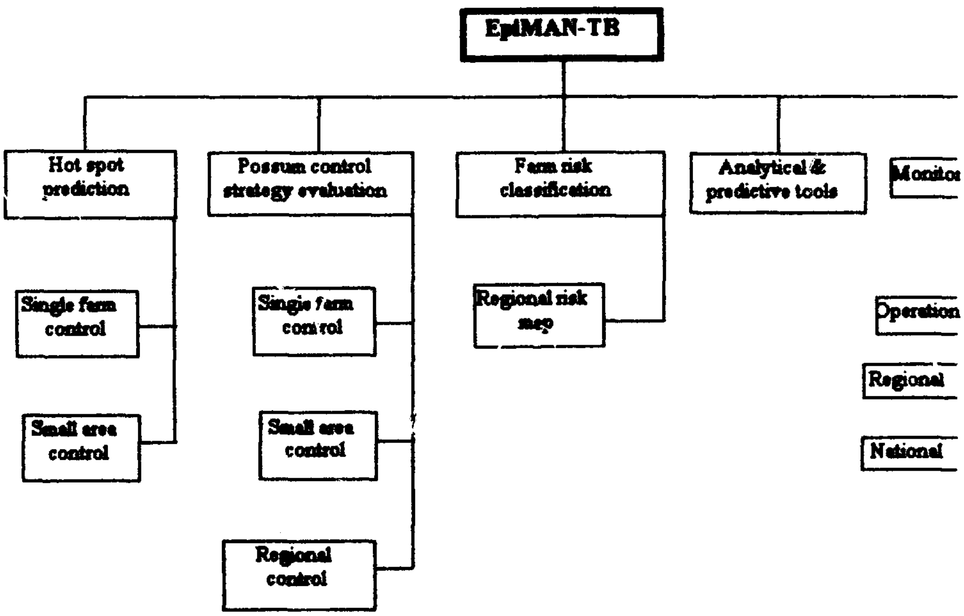


Figure 2. An outline of the tasks available within EpiMAN-TB

ity. Currently vegetation is divided into four main habitat categories: forest, scrub, bush/pasture margin, and pasture. These categories may be modified as research results identify different habitat types that are associated with the density of possums and/or the transmission of TB between possums. Vegetation maps are currently in IDRISI raster format with 20-meter pixels. However, maps with different pixel sizes and map formats can be incorporated into the model.

The basic geographic unit in PossPOP is a possum den site at a 1-meter point location. The vegetation map is used to "populate" the model with both possums and possum den sites on the farm or area of interest. The densities of each vary with the vegetation cover. For example, the density of possum dens on pasture is very low but is higher in scrub. PossPOP can also use the habitat risk map produced by the hot spot prediction model to adjust the probability of TB transmission between possums in accordance with the vegetation cover. This enables the creation of 'hot spots' within the model. It also enables habitat risk factors to be taken into consideration in the design of control programs. For example, a program with the same level of possum reduction over the entire farm can be compared with a program that has a higher and more frequent population reduction in high and medium risk patches of habitat compared to low risk areas. The relative effects of these strategies on the incidence of TB in the possum population can then be compared.

The model requires a vegetation map to run. As for the hot spot prediction model, the geographic boundaries of the vegetation map can be defined either by entering a farm identification number or by an interactive process. If the user wishes to include the habitat risk map in the model, its geographic boundaries can be defined in the same way. Parameters associated with possum control programs that can be manipulated include: percent reduction in population, frequency and duration of population reduction, location over which the population reduction is applied. The output provided by PossPOP includes possum population parameters, TB infection parameters, and location of 'infectious' den sites. An example of the TB prevalence and population size, from a run of the model for two years with a control program producing an 80% reduction in the population, implemented 6 months from the start date, is shown in figure 3. An example of the geographic output including habitat classes and the locations of dens used by TB possums at the end of the two year period is shown in figure 4.

ii) District or regional control

A further model will be included in EpiMAN-TB to model the spread of TB through possum populations distributed

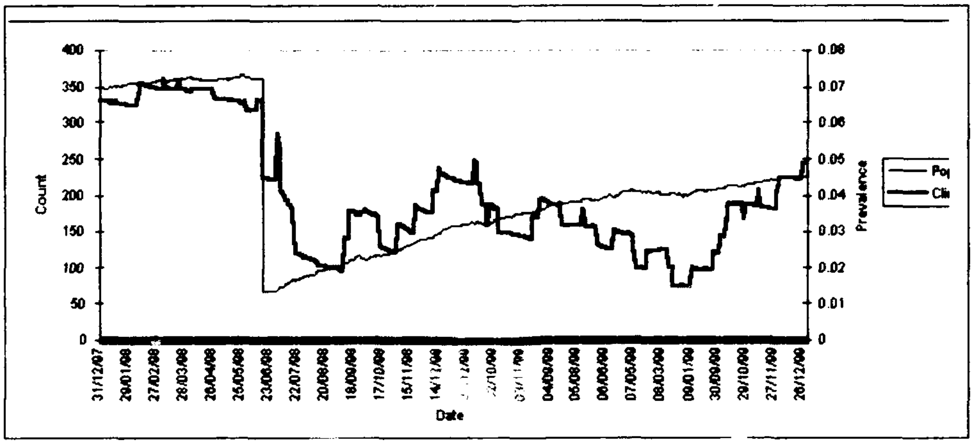


Figure 3 Graphical output from PossPOP, showing change in population size and prevalence of clinical TB over a two year period with the application of a control program in June 1998.

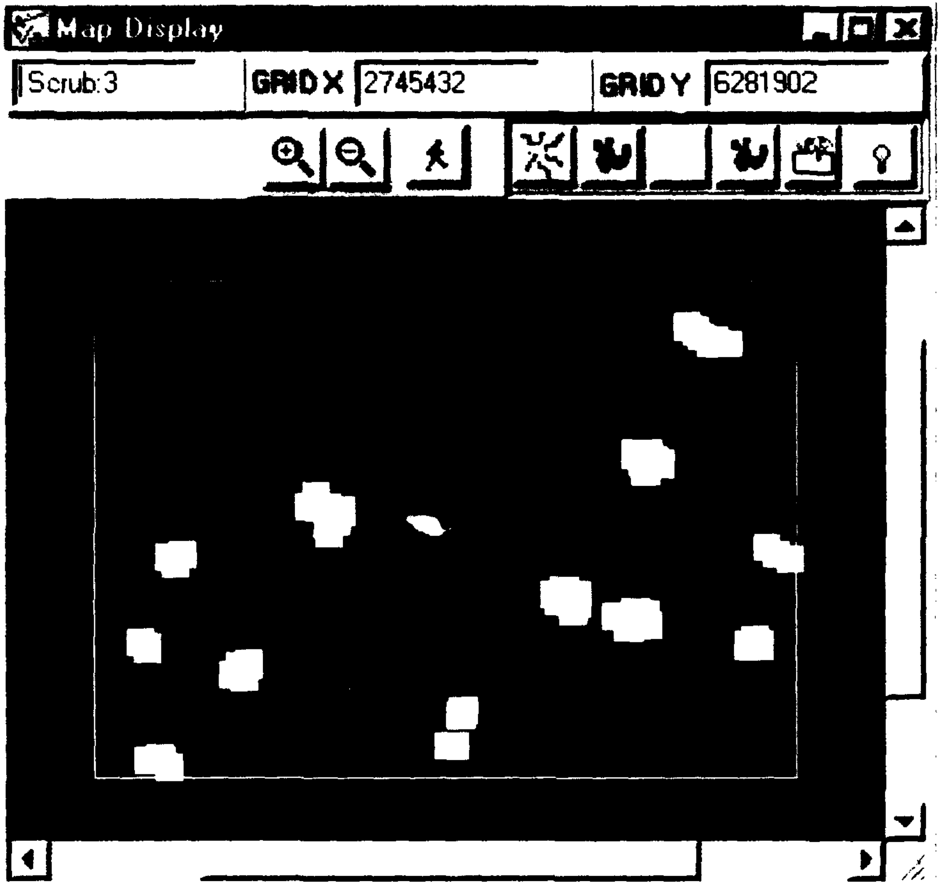


Figure 4 Snapshot of a vegetation map taken during a run of PossPOP, showing vegetation classes forest (black) and scrub (dark grey), and the location of infectious den sites (light grey)

over a larger area at district or regional level. The basic geographic unit will be a farm, and geographic boundaries such as major rivers and mountain ranges will be treated as semi-permeable barriers to the movements of possums. This model will enable the evaluation of possum control strategies implemented over the larger area, with farm units being populated by data derived from PossPOP, adjusted to reflect the circumstances of interest on the farm.

2.3.3 Farm TB risk classification

The ability to classify farms within a region according to the risk of TB in the on-farm cattle or deer population being high, medium or low would enable TB managers to differentiate the intensity with which control measures are applied according to the risk of the farm having a TB problem. This is particularly useful in an area where the possum population has recently become infected with TB as farms at the greatest risk of having infected possums on their property could be targetted more intensively for surveillance and disease control activities.

Current research is in progress to identify a set of geographic features of farms associated with high, medium and low levels of TB infection in the on-farm cattle. As for the hot spot prediction module, these factors will be used to generate an expert system to classify farms based on their

vegetation, topography and density of TB in cattle in their surrounding area.

2.4 Other tasks

Once development of the above components is complete, the software will be expanded to include other tasks which are considered useful in managing TB.

3. Conclusion

Results of research on the spatial distribution of TB in possums are now becoming available, providing information that can be used to predict the location of TB hot spots with a useful probability. At the same time farmers are being required to take greater responsibility for controlling the spread of TB on their farms. EpiMAN-TB provides tools that will assist TB field personnel working with farmers to develop specific programs for their farms. At the regional level EpiMAN-TB provides information that will assist the development of possum control strategies that focus control measures more tightly in areas where they produce the greatest effect. At national level EpiMAN-TB assists the making of policy decisions with respect to new control methods such as biological control of TB in possums and TB vaccination of possums. It also provides tools for the monitoring of disease control progress on a geographic basis across the country.

EpiMAN-TB is a comprehensive piece of software with easy access to the information required for the major decisions that need to be made with respect to the management of possum-associated TB in an area. This decision support system provides a way of integrating the current state of knowledge on TB and possums into disease management decisions, in the expectation that better decisions will be made.

Bibliography

Hickling G. (1995) Clustering of tuberculosis infection in brushtail possum populations: implications for epidemiological simulation models. In *Tuberculosis in wildlife and domestic animals*. F. Griffen and G. de Lisle (eds). University of Otago, Dunedin.

Pfeiffer D.U., Stern M.W., Morris R.S. (1994) PossPOP - a geographical simulation model of bovine tuberculosis infection in a wildlife population. Proceedings of the 7th International Symposium on Veterinary Epidemiology and Economics, 1994, Nairobi, Kenya. *The Kenya Veterinarian* (18) 2.

Using Spatial Similarity For Exploratory Spatial Data Analysis: Some Directions

A. Holt and G. L. Benwell
Department of Information Science and
Spatial Information Research Centre,
University of Otago, P. O. Box 56, Dunedin.
Aholt@commerce.otago.ac.nz

Presented at the second annual conference of GeoComputation '97 & SIRC '97, University of Otago, New Zealand, 26-29 August 1997

1 Abstract

Spatial information systems are now more fully incorporating artificial intelligence, mathematical and statistical modelling and other advanced analysis techniques for the extra computational analysis they allow. This is creating an increasing demand for computational power and the ability to handle very larger data sets. These demands and the associated solutions define the domain of geocomputation.

This paper further emphasizes the demand for geocomputational techniques by outlining an artificial intelligence technique called case-based reasoning. More specifically, this paper outlines a combination of case-based reasoning with spatial information systems and summarizes the computational techniques so derived to address spatial problems. The basic premise developed here is that spatial problems can be solved using similar spatial phenomena. Essentially a spatial problem is solved by searching a case base for another spatial case similar to the problem case. Then the knowledge/information obtained from the searched case is used to further an understanding of the phenomenon. This is a discussion paper outlining some directions for researching spatial similarity.

2 Introduction

Data exploring and data re-use techniques are set to having an increasing impact on information technologies. Case-based reasoning (Schank 1982), data mining and knowledge discovery (Fayyad 1997) are techniques used to search, recognize, extract, examine and predict decision knowledge from data. Earlier research (Holt 1996) has focused on applying case-based reasoning (CBR) techniques (in particular the *reuse* component of CBR) to spatial phenomena. The research direction is now focused on determining methods to store (represent) spatial data in a case and how this affects the *retrieval* component of CBR.

This paper details how cases are indexed for efficient retrieval and the similarity and weighting system between new and past cases. It is held that *spatial similarity* is an important concept for storing and retrieving cases. Spatial similarity will aid in determining clusters and feature detection for classification. This presupposes that it is possible to define spatial similarity. As a starting point, *spatial similarity is defined as those regions which, at a particular granularity (scale) and context (thematic properties) are considered similar.* Similar may be determined by any one of a number of methods - fuzzy membership (Lofti 1965), rough sets (Pawlak 1995) spatial auto-correlation and various statistical techniques to mention just a few. It is important to accept that similarity must be defined by variables that must be measured in some manner. Spatial similarity or spatial patterns would in turn help explain certain phenomena and their surrounding circumstances.

3 Research in Computational Methods for The Analysis of Spatial Data

Recent research has advanced the analytical capabilities of spatial information systems. Advances include, rule and knowledge-based approaches (Webster 1990; Smith & Yiang

1991), hybrid connection systems (Kasabov and Trifonov 1993) and a more innovative research approach where spatial reasoning is used to identify a given situation with other known typical scenarios (Williams 1995). AI-hybrids and other new strains (such soft computing, computational intelligence and linguistic representations of probability) may provide further possibilities.

Researchers outside the domain of spatial information systems have been focusing on similar themes and striving for a similar goal; understandably they approach the problem from different directions. Intelligent data analysis techniques for exploratory data analysis have accelerated the research into data mining and data trawling. Other examples are;

- geo-statisticians, working on spatial autocorrelation, fractals, spatial clustering, Kriging and anisotropy;
- AI scientists who are working on spatial representation, robotic vision, image analysis and processing.

4 *Case-Based Reasoning*

This section presents an overview of Case-based reasoning (CBR), the CBR cycle, and explains the main characteristics of the technology. CBR is a general paradigm for reasoning from experience. It assumes a memory model for representing, indexing and organising past cases and a process model for retrieving and modifying old cases and assimilating new ones (Kolodner 1993). The components of CBR, as extracted from the above definition, include representation, indexing and the storing of cases for problem solving by retrieving, adapting, explaining, critiquing and the interpreting of previous situations. This process is used to create an equitable solution to a novel problem using previous information. It is contended that these components be added to a spatial information system to complement its analytical functionality so as to build a spatial reasoning system.

CBR technologies can be used to:

- Solve a new problem for which a solution is unknown by retrieving and adapting similar problems that have been previously solved,
- Anticipate future events (decision support),
- Explore and analyse databases and generate hypotheses about the data.

CBR is adoptive because of its computational techniques and intuitive methodology. Jose et al. (1996) are using CBR to analyse remotely sensed images and assimilate spatial similarity by indexing and matching the vectors within the images. Smith et al. (1995) have developed a system called *Interactive Case-based Spatial Composition* which enables the user to interactively compose building layouts. CBR has been used for the better understanding of medical images (Grimnes and Aamodt 1996; Berger 1994) and meteorological images were the focus of research conducted by Jones and Roydhouse (1994) in trying to predict weather patterns. Their research focused on the efficient retrieval of structured spatial information. Goel et *al.* (1994) in trying to design robots that can navigate through space, used CBR techniques and a hierarchical spatial model for their experiment. Keller (1994) has conducted research combining GIS and CBR techniques. Keller used CBR for knowledge acquisition techniques for cartographic generalisation.

4.1 What Is A Case?

A case is the basis of a CBR system.

> *A case is a contextualised piece of knowledge representing an experience that teaches a lesson fundamental to achieving the goals of a reasoner* (Kolodner 1993:13).

Cases are experiences which have occurred. They are for example, problems that have been solved (or failed) by a problem solving mechanism (Althoff *et al.* 1994).

The major components (Kolodner 1993;Althoff *et al.* 1994) of a case include;

- *Problem/situation description:* the state of the real world at the time the case was happening and, if appropriate, what problem needed to be solved at that time,
- *Solution:* the stated or derived solution to the problem specified in the description or the reaction to its situation,
- *Outcome:* the resulting state of the world when the

solution was carried out,
- *Extensions:* context (justification); links to other cases; failures encountered.

4.2 The CBR Cycle

The basic steps in CBR are as follows (Aamodt & Plaza 1994).

- *Retrieving a past case (a problem and a solution) that resembles the current problem.* Past cases reside in the case base (memory). The case base is similar to a database that contains rich descriptions of prior cases stored as units. Retrieving a past case, involves searching for one which is similar and calculating its degrees of similarity which determines what features of a problem should be considered,
- *Adapting the past solution to the current situation.* Although the past case is similar to the current one it may not be identical. If not, the past solution may then be adjusted to explain differences between both problems,
- *Applying the adapted solution and evaluating the results,*
- *Updating the case base.* If the adapted solution is appropriate then a new case can be formed. The new case is composed of the original (or similar) solution and the repaired solution. It is stored in the case base so the new solution will be available for retrieval during future problem solving. In this way, the system becomes more competent as it gains experience.

5 *The Spatial Reasoning System*

This section outlines the concept of a spatial-AI-hybrid called the spatial reasoning system (SRS). It is presently under research and development. The concept has arisen from the belief that GIS are limited in reasoning ability and CBR can be integrated to support this deficiency. The primary use of such a system will be to develop reasoning techniques for discovering knowledge about areas which are considered to be *spatially similar*. CBR offers: the ability to reason, explanation features, adaptation facilities, extended generalisation techniques, inference making abilities, constraining a search to the solution template, solution generation, the ability to validate and maintain knowledge bases. These features would aid planning, forecasting, diagnosis, design, decision making, problem solving and interpretation.

Some definitions of spatial reasoning will include spatial cognition and the representation of knowledge (Hernandez 1993; Williams 1995). Frank (1996) defined reasoning as "*the conceptualisation of situations as space*". Others define the term to mean the ability to reason, learn, think and to draw conclusions from facts (Holt 1996). The latter is preferred here, though the term *spatial discovery* is gaining in popularity and may well be an appropriate compromise.

The SRS will eventually be used;

- As a problem solving tool which has the ability to reuse previous similar spatial problems and their solutions to solve a current problem (Holt 1996),
- As an exploratory spatial data analysis technique for data mining/trawling and pattern searching/matching,
- As an alternative method to represent and store spatial data. Storing data as spatial cases, equivalent to object oriented languages, but having the added benefit of learning features.

The following are examples of questions the SRS may address;

- *The SRS hybrid is used to facilitate searches and solve the following problems:*
  Are there spatial phenomenon similar to the searched example? Identify unique areas, evidence of trends, patterns or other variations. If so, what attributes are associated with that phenomenon? In finding a similar spatial pattern a GIS may be to display and store data. CBR provides the functionality to find a similar pattern and, more importantly, to analyse its properties. These properties would extend from the obvious spatial pattern to other attributes associated with the pattern. This functionality could be used for classification or in solving complex problems using previous experiences.
- *Providing new opportunities in spatial analysis via information retrieval and pattern recognition. To solve the following*

problems;

Is there evidence of clustering with respect to specified sources or possible causes? What spatial associations exist between cases?

- *The SRS hybrid is used to facilitate queries and solve the following problems:*

  Which spatial phenomena have the certain criteria? What attributes are associated with a spatial phenomenon with these criteria?

5.1 Why Use The *SRS*?

CBR offers the potential for improved functionality to current GIS. This is achieved in a complementary fashion as the functions they both have are executed in different methods (for example, retrieve and retain). The functions of GIS and CBR techniques which differ the most are their abilities and techniques for representing and storing data. The ability of CBR to learn is another component which separates it from a GIS. Data and knowledge in the form of cases are stored and represented so they can be retrieved quickly to suit particular requirements. This complicated storing method, existing of bundles of knowledge, is indexed to allow new experiences to be saved. A sense of learning, therefore, is introduced. Other components offered by CBR include the reuse and revise (adapt) functions which current GIS software packages lack.

6 *The Processes Within A CBR System*

The sequence for running a CRB sub-operation within a conceptual SRS would be as follows;

- The user provides a case for comparison,
- The program performs an index search and finds a subset of cases that match all the index constraints. The index constraints are taken from the field values provided by the user. The program searches the case base for the subset of cases that match all the index constraints exactly,
- If no cases match all the index constraints (for instances when there are only a few cases in the case base), the system prompts the user to search for different index values. If there are no cases which match all the index constraints, the user is informed and is prompted to enter new values for the index constraints. These may be made more general by specifying abstraction values or by specifying fewer constraints,
- A case is selected from the subset. After the index search is completed the case matcher is invoked to scan the subset of cases to find the one with the highest weight value. This is selected and the repair rules are then applied,
- Repairs are carried out on the selected case. On occasions additional information is requested after a case has been selected. Sometimes a repair rule can cause the current case to be abandoned and the selection process to begin again,
- If the user is dissatisfied with the previous matching case(s) further cases may be examined. This is continued until they are satisfied with a matched case or until the user exhausts all possibilities.

Case file blocks of code are required to define; *introduction, case definition, index definition, modification definition, weight rule definition, repair rule definition, and case instance*. The *introduction* block contains introductory text which is displayed when the program has finished checking the case file. The *case definition* sets the types and the weights of the problem fields that may appear in a case. The information in the *case definition* is used for checking input cases while the weights are used to aid the case-matching process. The *index definition* sets the fields used as indexes when searching for a matching case. A case base should have at least one field used as an index. The type of index field must be enumerated. The *weight rules definition* sets rules that may be applied to change the weights used for matching cases. The *modification definition* sets the modification rules and provides a means of specifying that certain symbols or numbers are similar. This is undertaken first for matching purposes and provides a means of specifying symbols as abstractions of others and second for making the search more general or for defining generalised cases. The *repair rule* definition contains the repair rules. These are used to modify the solution retrieved from the case-base making it more suitable for the current situ-

ation. Both the *modification definition* and the *repair rule definition* may be omitted. To be a complete CBR system, however, it should contain both. The last set of blocks are the case instances. These make up the case base. The case file must contain at least one case instance and will initially need to be seeded with many cases before it is operable.

6.1 Case Matching

The components that will be used to evaluate the spatial similarity of cases include case matching and case retrieval.

> *Case retrieval, no matter the method, requires a combination of search and matching* (Kolodner 1993)

Organisational structures are searched to find potential matching cases, and each is evaluated for its potential usefulness. The evaluation is done by matching functions. It is necessary before discussing retrieval, however, to discuss the fundamentals of matching. The following three stages are indicative of the case matching process. These include weight rules, index values and weight-matching;

- Weight rules may be applied to find a set of appropriate weights for performing case-matching.
- The index values, which are either taken from the user case or specified separately by the user, are used to perform an index search. This retrieves a subset of cases from the case base which match all the index values exactly (except when abstraction symbols are specified as index values, in the modification rules).
- Once this list of cases has been retrieved the user can allow the program to automatically select a case. This is based on weight-matching. For each case in the subset the case-matcher finds a weight which is obtained by totaling the weights of all matched fields. Fields which do not match exactly, but are defined to be similar by the modification rules, return a value which is less than the field's normal weight. The case-matcher selects the case with the highest total weight. The user can browse through the selected cases and select a case manually.

The method of case-matching consists of two phases. First, the enumeration-type fields cited in the index block are used to select a sub-set of cases from the case-base. Second, a form of nearest-neighbour matching is used to select the best case from the subset. The weights are not attached to the cases themselves. The system parses through each case in the subset evaluating their weight. A record of the best matching cases are recorded. The importance of each field is defined in the case definition section. Internal rules (not to be confused with the weight rules block) are used to evaluate what proportion of the weight is returned for each field. If for example, the values match exactly then the full weight is returned. In comparison, if two enumeration symbols are similar then 0.75 of the field weight is returned. Strings have to match exactly or zero field weight is returned. During the parsing, of two lists of symbols and for example, if half of them match, then half of the field weight is returned.

After the case has been selected extra values may be input if the case has a local field definition associated with it and then the global repair rules (in the *repair rule definition*) are actioned. Furthermore, if there are *local repair rules* associated with the case then these will be actioned. If a repair rule causes a re-selection to occur, another case is selected using weight matching and local fields may again need to be entered. The repaired case is displayed and the user is given the option of adding the repaired case to the case base. If the user adds the repaired case to the case base, the values for the name and result of the case, which is then appended to the case file and added to the case base.

6.2 Retrieval Methods

Given a description of a problem, a retrieval algorithm, then using the indices in the case base should retrieve the most similar case(s) to the novel problem or situation. The retrieval algorithm relies on indices and the organisation of the memory to direct the search to potentially useful case(s). Methods for case retrieval include, bounds-test for nearest neighbour search, induction, knowledge guided induction, structuring using the inter-quartile distance, the k-d tree, similarity measuring in the k-d tree, exemplary 2-d tree, template retrieval, discrimination net-

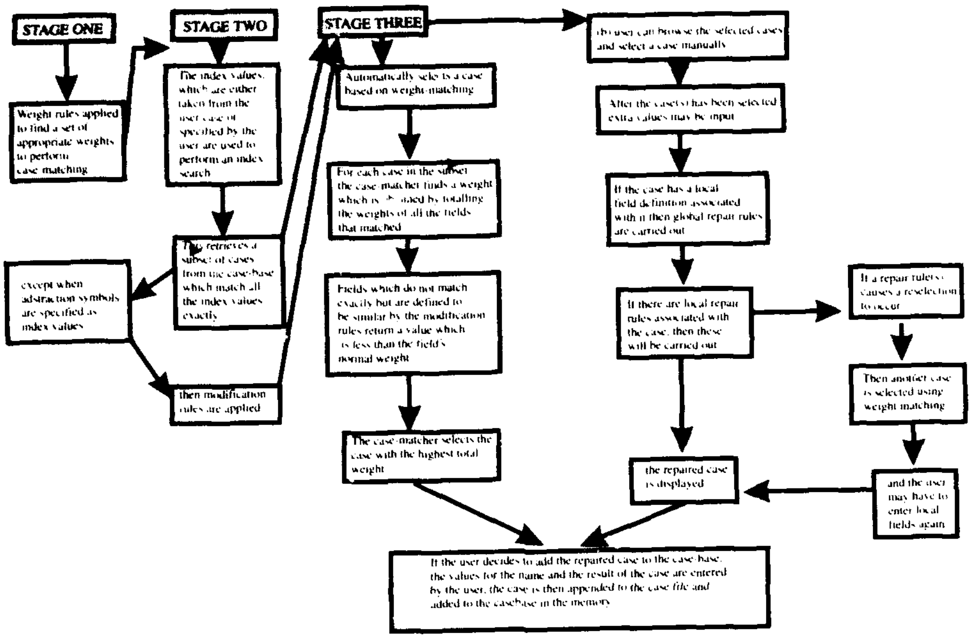


Figure 1 The Stages Involved in the Case Matching Process

works and parallel retrieval (Leake 1995). These methods can be used alone or combined into hybrid retrieval strategies (Althoff *et al.* 1994).

The retrieve task (shown as the shaded parts of figure 2) starts with a new case description and ends when a best matching previous case has been found. Its sub-tasks (also shown as the shaded parts of figure 2) are called *identify features, initially match, search* and *select* which are executed in that order. The identification task produces a set of relevant problem descriptors, the goal of the matching task is to return a set of cases sufficiently similar to the new case, and the selection task works on this set of cases and chooses the best match. Some case based approaches retrieve a previous case, based on superficial, syntactical similarities among problem descriptors, while other approaches retrieve cases based on features that have deeper, semantic similarities. To match cases based on semantic similarities and relative importance of features, an extensive body of general domain knowledge is needed to produce an explanation of why two cases match and how strong the match is. Syntactic similarity assessment (a knowledge poor approach) has its advantage in domains where general domain knowledge is difficult or impossible to acquire. Conversely, semantic oriented approaches (knowledge intensive) can use the contextual meaning of a problem description in its matching for domains where general domain knowledge is available (Aamodt & Plaza, 1994; Althoff, *et al.* 1994; Zeleznikow 1995).

The retrieval function is also executed differently by the two systems. Zeleznikow (1995) suggests that the retrieval for CBR involves characterising the input problem by assigning appropriate features to it, retrieving the cases from memory with those features and similarity assessment. Similarity assessment determines the level of match between old and new cases (Althoff *et al.* 1994). CBR uses a nearest neighbour technique (indexed and case matching) for retrieval and a GIS system generally uses a relational database technique, making it not as flexible as CBR. For example, using structured query language to select a line which has the length of 23.5km or select a line where length is 20 to 25km. Whereas, CBR has *fuzzy* matching

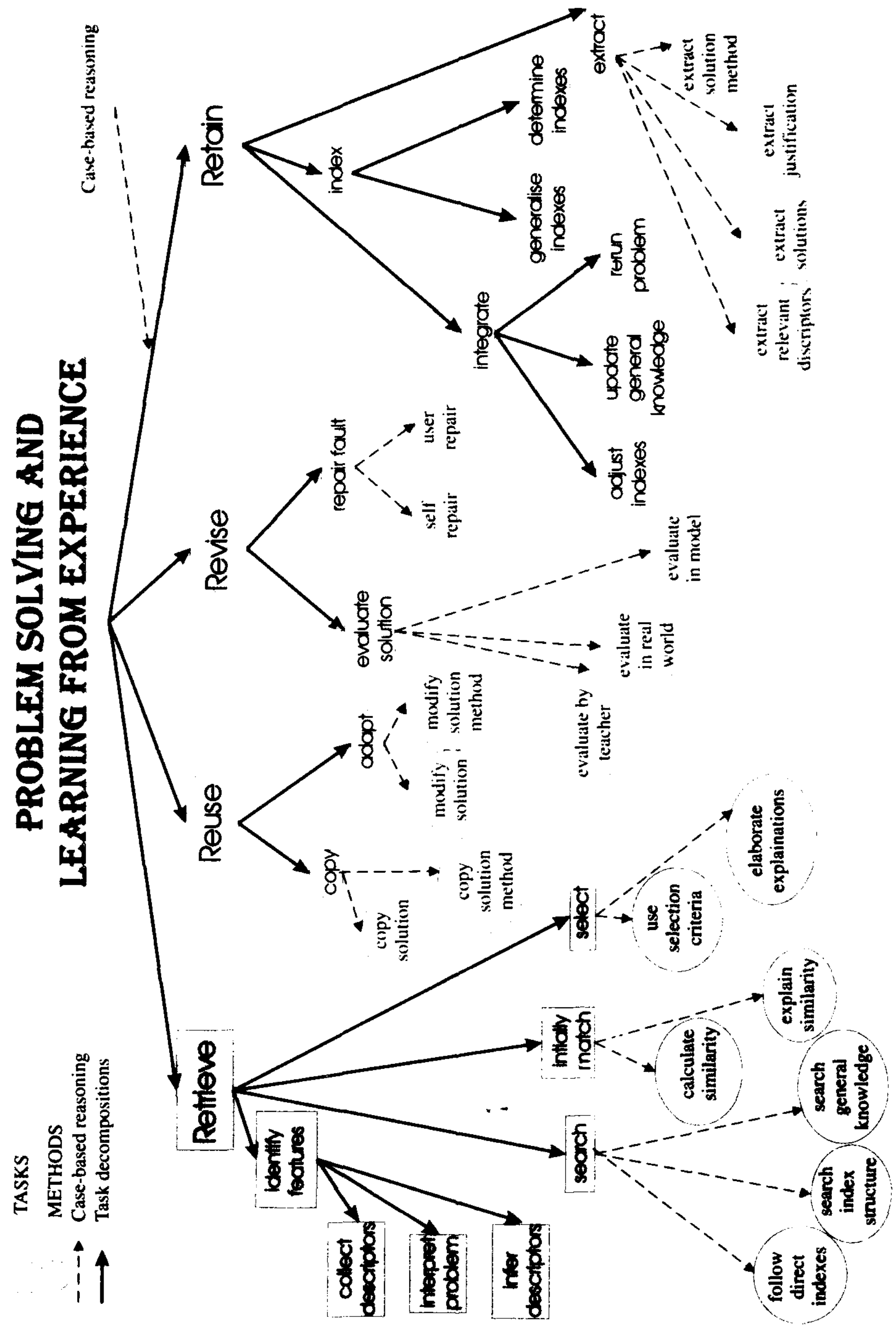


Figure 2. A Task-Method View of CBR (Aamodt & Plaza 1994)

which is achieved using the case and the weight matcher. For example, selecting a case where the line length equals 23.5 km. CBR will find solutions displaying all line(s) with a similar length to 23.5km together with any other information associated with that line.

Case indexing involves assigning indices to cases to facilitate their retrieval and is vital in identifying the most similar previous case(s). This involves the identification of the relevant factors in a case upon which the case based retrieval system can index the cases. Choosing indices manually involves deciding a case's purpose with respect to the aims of the reasoner and deciding under what circumstances the case will be useful. Several guidelines on indexing have been proposed by Watson (1994) stating that indices should;

- address the purposes the case will be used for,
- be abstract enough to allow for widening the future use of the case base,
- be concrete enough to be recognised in the future,
- be predictive of important case features.

7 Future Research

Similarity functions, case structure, domain data, reusability and problem solvability are some components which affect the similarity result and are being researched. Spatial case representation, multiple case representation, constraints (ensuring that an answer must be chosen from a particular data set), features (imagine this in a meta-data sense), goals and formalisms (representation formalisms in CBR and GIS) all affect the different retrieval types and the hence affecting the search for the spatial similarity . Experiments with these issues, will form the basis of the spatial discoveries, the new direction for research.

It is acknowledged that if large spatial databases are to be searched then improved sampling techniques must be used. CBR may just do that by indexing fields, or using template retrieval, induction algorithms or knowledge guided induction. Template retrieval, is similar to SQL-like queries. Template retrieval returns all cases that fit within certain parameters. This technique is often used before other techniques, such as nearest neighbour, to limit the search space to a relevant section of the case-base. Induction algorithms determine which features do the best job in discriminating cases, and generate a decision tree type structure to organise the cases in memory. This approach is useful when a single case feature is required as a solution, and where the case feature is dependent upon others. Knowledge guided induction, applies knowledge to the induction process by manually identifying case features that affect the primary case feature. This approach is frequently used in conjunction with other techniques, because the explanatory knowledge is not always readily available for large case bases.

8 Conclusion

In searching for a richer data model for encoding, searching and comparing complex geographical entities this paper has outlined a method that may allow more advanced analytical techniques to be executed on the geographical entities. This paper has proposed some possible directions to advance current GIS techniques for analysing, searching, recognising and extracting information on spatial patterns. This paper has outlined how an AI technique called case-based reasoning could help in achieving these proposed advances.

References

Aamodt, A. and E. Plaza, (1994) Case-based Reasoning: Foundational Issues, Methodological Variations and System Approaches. *Artificial Intelligence Communications*, Vol.7 No.1.

Althoff, K. Wess, S. and M. Manago, (1994) Reasoning With Cases-Theory and Practice ECAI.

Berger, J. (1994) Roentgen: Radiation Therapy and Case-based Reasoning. *In the Proceedings of the Conference on Artificial Intelligence Applications*, Proceedings. pp.171-177.

Burrough, P.A. and A. U. Frank, (1995) Concepts and Paradigms In Spatial Information: are Current Geographic Information Systems Truely Generic? *International Journal of Geographical Information Systems*. Vol.9.

No.2. pp.101-116.

Fayyad, U. M. (1997) Editorial. *Data Mining and Knowledge Discovery*, 1(1): 5-10.

Fischer, M. M. and P. Nijkamp, (1993) Design and Use of Geographic Information Systems and Spatial Models. In: *Geographical information systems, spatial modelling, and policy evaluation*, Eds. M. M. Fischer and N. P. Springer Verlag, Berlin.

Frank, A. U. (1996) Qualitative Spatial Reasoning: Cardinal Directions as an Example *International Journal of Geographical Information Systems*, Vol.10. No.3. pp.269-290.

Goel, A. K. Ali, K. S. Donnellan, M.W. Garza, A. G. D. S. and T. J. Callantine. (1994) Multistrategy Adaptive Path Planning. *IEEE Expert*, 9(6): 57-65.

Grimnes, M. and A. Aamodt. (1996) A two layer case-based reasoning architecture for medical image understanding. *EWCBR 96*.

Hernandez, D. (1993) Reasoning with Qualitative Representations: Exploiting the Structure of Space. *In Piera Carrete and Singh*. pages 493-502.

Holt, A. (1996) Incorporating A New Computational Reasoning Approach to Spatial Modelling. *1st International Conference on GeoComputation*, University of Leeds, Leeds, England. Vol 1. pp 427-442.

Holt, A. and G. L. Benwell, (1996) Case-based Reasoning and Spatial Analysis. *Journal of the Urban and Regional Information Systems Association*, Vol.8. No.1. pp 27-36.

Holt, A. and G. L. Benwell, (1995) Spatial Reckoning using Case-based Reasoning. *8th Australian Artificial Intelligence Conference Incorporating the Workshop on AI in the environment*, Proceedings. Canberra, pp. 53-66.

Jones, E. K. and A. Roydhouse, (1994) Spatial Representations of Meteorological Data for Intelligent Retrieval. *The Sixth Annual Colloquium of the Spatial Research Centre*, Proceedings. Eds. G.L. Benwell and N.C. Sutherland. Dunedin, New Zealand. pp.45-58.

Jose, B. Singh, B. P. Venkataraman, S. and R. Krishamn (1996) Vector Based Image matching for indexing in Case Based Reasoning Systems. *4th German Workshop on CBR*. pp 88-94.

Kasabov, N. and R. L. Trifonov, (1993) Using hybrid Connectionist Systems for Spatial Information Processing. *The Fifth Annual Colloquium of the Spatial Research Centre*, Proceedings. Eds. Benwell G. L. and N. C. Sutherland. Dunedin, New Zealand.pp85-96.

Keller, S. (1994) On The Use of Case-based Reasoning in Generalization. *Advances in GIS Research, Proceedings of the 6th international Symposium on Spatial Data Handling*. Eds Waugh and Healley, Vol 2:1118-1132.

Kolodner, J. (1993) Case-Based Reasoning. San Mateo, Morgan Kaufmann Publishers.

Laurini, R. and D. Thompson, (1992) Spatial Knowledge. In: *Fundamentals of Spatial Information Systems*, Academic Press, London.

Leake D. B. (1995) Case-Base Reasoning: Issues, Methods and Technology, *A Tutorial for the First International Conference on Case-Based Reasoning*, Sesimbra, Portugal.

Leung, Y. and K. S. Leung, (1993) An Intelligent Expert System Shell for Knowledge-based Geographical Information Systems: 1. The Tools. *International Journal of Geographical Information Systems*, Vol.7. No.3. pp.189-199.

Pawlak, Z., Grzymala-Busse, J. W., Slowinski, R. and W. Ziarko, (1995) Rough Sets. CACM 38(11): 88-95.

Schank, R. (1982) Dynamic memory: A theory of leaning in computers and people. Cambridge University press. New York.

Smith, I. Stalker, R. And C. Lottaz. (1995) Interactive Case-based Spatial Composition. *ICCBR-95*. pp20.

Smith, T. R. and Yiang, J.E. (1991) Knowledge-based Approaches in GIS. In: *Geographic Information Systems*, Eds. D.J. Maguire, M. F. Goodchild and D. Rhind. Longman

Scientific and Technical, Essex.

Watson, I.D. (1994) The Case for Case-Based Reasoning. In, Proc. *Information Technology Awareness in Engineering Conference*, 21-22 November 1994, London.

Webster, C. (1990) Rule-based Spatial Search. *International Journal of Geographical Information Systems*, Vol.4. No.3. pp.241-259.

Williams, G. J. (1995) Templates for Spatial Reasoning in Responsive Geographic Information Systems. *International Journal of Geographical Information Systems*, Vol.9. No.2. pp.117-13.

Zadeh, L. (1965) Fuzzy sets. *Information and Control*, Vol.8. pp 338-353.

Zeleznikow, J. (1995) Merging Symbolic Reasoning with Subsymbolic tasks. A Tutorial at the 8th *Australian Joint Artificial intelligence Conference*, Canberra, pp147.

Data Questions in GeoComputation

Brian G Lees
Department of Geography,
Australian National University
ACT 0200,
e-mail Brian.Lees@anu.edu.au

AIR NEW ZEALAND
Air New Zealand Guest Keynote Speaker

Presented at the second annual conference of GeoComputation '97 & SIRC '97, University of Otago, New Zealand, 26-29 August 1997

1: Introduction

The new computational solutions which have developed in parallel with the widespread uptake of computer power in geography tend to represent a distinct move away from the more traditional, parametric statistical, methods introduced to the discipline during the so-called quantitative revolution. The implications of adopting these new methods have not yet been fully appreciated by many researchers, indeed the retreat by large sections of the discipline from any serious engagement with quantitative approaches to geographical problems means that there is a growing potential for misuse and abuse of these solutions as they become more accessible.

It is well worth remembering that our new computational solutions comprise several critical components. Firstly, the new hardware configurations without which they could not be implemented. Secondly, the new algorithms themselves. Thirdly, the data, and fourthly, the problem. Perhaps this fourth element could be better described as the problem statement. Success depends on the adequacy of all of these, and on their correct integration. The data, the problem statement and the links between the algorithm and the data, between the data and the problem statement, and the problem statement and the algorithm have all received much less attention than the matching of algorithms and hardware. As we become increasingly engaged in harnessing our new computer power to geocomputation it is perhaps worth casting our eyes over these other, equally important, facets of quantitative investigation, analysis and prediction in the geosciences.

In this discussion I will refer to only two of the large suite of data-driven modelling techniques rather than range across the field. There is simply neither time, nor space, to do otherwise. Most of the points I want to make are relevant to a much wider group of techniques and these are merely convenient examples for discussion. The two I will use are decision trees and so-called back-propagation. Both travel under fancier names on occasion, but these labels are broadly understood and will suffice.

2: Data

Any examination of data in geocomputation needs to consider at least the data distribution, the data model and the way in which the data has been sampled. The precision and accuracy of the data, both in the spatial and attribute domains, also need to be considered. These latter aspects of data have been extensively dealt with elsewhere (Goodchild & Gopal, 1989).

2.1: Data Distribution

After decades of accepting the questionable proposition that most phenomena in the natural world are normally distributed we are now adopting non-parametric methodologies with enthusiasm. The fact that many of these can be parallelised has, perhaps, added to this enthusiasm. There are some costs in this enthusiasm. It is safe to say that, given a data distribution approximating a normal distribution, parametric methods tend to produce a better result than non-parametric methods unless large, or carefully chosen, samples are selected.

To produce results of matching quality in the analysis of non-normally distributed data using non-parametric methods, much larger and more carefully structured samples are needed. By definition, non-parametric, supervised inductive learning systems have no information on the distribution of the data other than that which can be inferred from the learning sample. Few of the studies which have appeared in the literature indicate that this has been recognised. That this is such a problem is perhaps due to the fact that many of those involved in this branch of geocomputation are not data gatherers, but data processors. There is a real temptation to use 'legacy' data sets for experiments in geocomputation and this leads to the use of proportional samples. Given the error minimisation rule on which many of these systems are based, the use of a proportional sample as a learning sample will bias the system towards the largest categories.

2.2: Data Models

Much of the published work on data models focuses on the data model as the rationale for organising data in the computer. In computer science it is a means of capturing the semantics of the data through definitions of the operations related to classes, describing which combinations of operations are legal, which combinations of operations are equivalent, and consistency constraints among data. This bias towards the computer science view of data models is quite understandable as it is a necessary tool to deal with the data, but many phenomena have not been carefully scrutinised by domain experts in the same way and I suspect that, when this happens, the whole concept of data model will become considerably more complex and critical.

When we start to consider whether the measure used to code the data is appropriate to the phenomena we wish to examine we need to remember that many disciplines, including geography, routinely classify data as part of their collection protocols. This pre-analysis processing is often not recognised as such but can be a major limitation to accurate prediction based on such sampling. All too often phenomena distributed as a continuum are discretised into gaussians on the assumption that this is an appropriate data model for the phenomenon. The type of measure used is also critical. The use of nominal measures, rather than ratio or interval measures, increases the requirement for an unbiased sample significantly. Whilst ordinal measures are not as difficult to deal with as nominal measures, they are considerably less informative than, say, an interval measure.

The problem outlined above is the natural consequence of a habit widespread through many disciplines. The classification of data prior to analysis is almost an unconscious act for many field scientists. That this is unnecessary now that we are no longer bound by the cartographic model of spatial data has not really penetrated the consciousness, and standard procedures, of many disciplines. Indeed, in many cases, the data collection itself imposes this structure. The step between observation, and the recording of that observation is often one in which some form of classification takes place. The value of each observation, as a unique data point, is then immediately degraded.

All other things being equal, if one can provide a learning system with some indication of how values in an attribute relate one to another, then the system will do a better job. Humans like to simplify these relationships as we are unable to deal very effectively with high frequency variability in data. By coding data to suit human perceptions, we degrade it and remove information a non-human learning system may be able to interpret. For example, in many natural systems tasks geology is an important variable. The taxonomy in geology being what it is, the relationship between a granite, an Essexite and a Monzonite, and the lack of a close relationship between those and a Sandstone are not apparent (to an algorithm) from the class numbers used to represent these in a GIS. It is necessary to recode these categories using some appropriate interval or ordinal scale, in the case of an erosion study 'K' values would be appropriate. The 'K' value is a ratio value with a direct relationship to erodibilty. In the case of vegetation modelling, geology can be recoded according to some interval or ordinal scale of nutrient status. Deriving appropriate

measures requires a knowledge of both the attribute, and the interactions of attributes relating to the phenomenon being modelled.

These simple pre-processing stages are needed overcome the knowledge gap which exists between human and algorithmic 'intelligence'. Most natural scientists understand the relative difference in nutrient status between weathered granite and sandstone. The names communicate a suite of attributes to the expert human listener. Unfortunately, there is no inherent information in the terminology to inform either the non-expert human or algorithm. Even worse, because of the necessity of labelling attribute classes with numerical identifiers when data is imported to a GIS, there is sometimes a tendency to carry out analyses which improperly utilise the mathematical relationships between identifiers, when no such relationship is implied. This is a common trap for non-expert users, but it is also a trap for expert users working with data from domains in which they are not expert.

2.3: Data Sampling

I have already mentioned the importance of sample characteristics briefly. In the use of optimising, or error minimisation, techniques, it is important that each case one wishes to predict or classify is equally well represented in the learning sample. Proportional sampling techniques will not produce this. One must resort to quite structured, stratified methods to achieve this sort of sample. One must also attend closely to the scale at which one samples. Now that we can move away from the restrictions of the cartographic model, many disciplines have not yet understood that data scale and display scale are no longer synonymous and need to be considered separately. For our purposes, the display scale is much less important than the scale at which the data was measured. This is particularly true when one is looking at context, spatial or temporal.

Both spatial and temporal variability are strongly scale dependant. There is a general trend in most land cover data for spatial autocorrelation to be low at fine scale, to rise to a maximum at an intermediate scale and then to decline. One can see a similar pattern in many forms of temporal data. The diurnal range of bio-activity, illumination, temperature and pressure is often nearly as great as the annual range (based on daily observations), and much greater than the inter-annual range. We need to move to epochal time scales to see the diurnal range exceeded. We filter out the fine scale variations when we make observations, but we tend do this informally. To reduce data based error in analyses it is important that we exercise more conscious control of input data scale. If we cannot control it, then we need to be aware of the consequent errors.

Spatial and temporal variability also depends on the data space, or domain, in which one views the data. Spatial data exists in a number of discrete domains (Lees, 1994; Aspinall & Lees, 1995). In each of these there exist topological relationships, but these relationships vary from domain to domain. We are most familiar with spatial data existing in a geographic space defined by latitude, longitude and elevation. Movement from point to point in this space is a vector. It is not possible to move from one point to another without transiting intermediate points. Each point is unique.

In the other, conceptual, domains or data spaces topological relationships are different. These data spaces can be spectral space, environmental data space, even socio-economic data space. The fundamental, and shared, characteristic of these spaces is that movement through the space has a logical meaning. Spectral space, for example, forms the basis for most analysis of remotely sensed data. Proximity suggests similar colour. Trajectories of reflectance values for developing crops on different soils form the basis for the common Kauth-Thomas, or Tasseled Cap, transformation. Trajectories in spectral space form the basis for sub-pixel modelling of vegetation structure. In these analyses vectors represent changes in the reflectance at a point, through time. No motion in geographic space is envisioned. A large number of points in geographic space can occupy a single location in spectral space. The converse is not true.

In environmental data space, the basis for environmental domain analysis, topological relationships are linked directly to environmental gradients. Vectors in this space drive the continuum of change in vegetation composition observed in nature. The conflict in ecological literature between those who favour a community view of vegetation and those who view it as a continuum lies squarely on the fact that community is a spatial concept in geographic space, whilst the continuum is a spatial concept in environmental data space (Austin and Smith, 1989). Both are common representations, but fundamentally different in the way they can be analysed. In geographic space one can move from one point to another along a vector. This same motion in environmental data space may result in no motion, if the environments along this vector in geographic space are the same, or a jump from point to point if say, a soil boundary is crossed. As before, a large number of points in geographic space can occupy a single location in environmental data space and, once again, the converse is not true.

This particular dichotemy, between representation of vegetation distribution in geographic space and environmental data space, is a dichotomy between data models. The 'mapping' school reduce observations of vegetation to a series of vegetation classes, even forest types. In some ecosystems, particularly Australian eucalypt forests, these class boundaries are cultural (statistical) artefacts. Slight changes in contribution to the canopy can lead to a change in class. In such cases, there is often more variation within the class than between classes. Nevertheless, the fundamental structure of choropleth mapping requires this reduction of variance to permit the mapping of polygons. This mismatch between the phenomenology of the data and the data model, excusable in the days where choropleth mapping was the only means of representation, has been carried forward to the present.

Domain knowledge is fundamental to constructing the necessary spaces for analysis, and for understanding the relationships between the spaces. In many problems different parts of the analysis need to be carried out in different data spaces. Importantly, a sample which can be considered to be representative in one domain may not be representative in another. Sampling strategies therefore need to consider the data distributions in all of the relevant domains.

3: Interactions Between the Algorithm and Data

In parametric statistics a classifier is an algorithm, in non-parametric, data-driven analyses the classifier results from the interaction between an algorithm and a learning sample. The characteristics of the learning sample determine, to a large degree, the behaviour of the classifier. Careful design of learning samples is vital for good performance in this area. The behaviours of the different algorithms in the way they use the learning sample is also very important in the design of analyses.

3.1: Decision Trees

The recursive partitioning which is the basis for decision tree algorithms seemed to be an ideal strategy for dealing with the data domain problem. Each split, or decision rule, is made in only one data domain. The tree building (learning) procedure moves from data domain to data domain as it searches for optimum splits and makes only minimal assumptions about the relationships between variables. This sort of inductive learning produces clear and explicit results. Careful monitoring of the derived rules is necessary to identify rules based on statistical artefacts rather than process relationships. This monitoring, preferably by a domain expert, is vital to weed out nonsensical relationships which would induce error when the tree was used as a classifier. High correlations between independent variables often confuse this sort of system. For example, in the modelling of vegetation distribution around Kioloa a decision tree may indicate that elevation is an important variable. Examination of the tree will show that geology is an alternate split at that point. The high correlation between geology and elevation in the Kioloa learning set is a statistical artefact of the data set. The area is predominantly Sydney Basin sediments which are flat lying. Changes in geology correlate with changes in elevation for much of

the data set and the digital elevation model is the higher resolution variable. It therefore comes up as being more significantly related to change in tree species than does geology. However, as the elevations concerned are not extreme enough to generate significant climatic gradients, it is clear that the process driving the change in species is the slight change in nutrient status associated with the different geology types. A domain expert would be able to identify this quite readily and change the variable at that point accordingly. Slope also acts as a useful correlate for changes in geology, often at scales well below that at which geological information is available. This explicit nature of decision trees is very attractive but in many applications does not offset their hunger for huge learning samples.

3.2: Artificial Neural Nets

Having experimented with the decision tree approach for some time with good results (Moore et al., 1991; Lees & Ritman, 1991) it became clear that, for some applications, the amount of learning data required to produce the required level of discrimination (number of classes) was impractically high. This is particularly true where some classes are poorly represented in the learning sample as, with a stopping point of 25 or 30 points, many classes simply have no chance of being predicted. It is possible to plot probability surfaces, or fuzzy set membership, using the membership of the populations at each terminal node to overcome this, but these problems prompted a further search for methods less hungry for data. After a short search, several types of Artificial Neural Net appeared to offer attractive solutions to the problem (Fitzgerald & Lees, 1993; 1994). It is useful to think of some of these algorithms as doing in parallel what decision trees do in series.

Artificial Neural Nets are a field, rather than a group, of quite unrelated algorithms. Many originated as projects to understand human information processing and were never intended as the analytical tools they are now sometimes seen as being. Neural Nets are part of a suite of data-driven modelling techniques which are useful when the processes underlying a phenomenon are either unknown, only partially known, or would necessitate the generation of an impracticable level (scale, volume or cost) of input data. Within the suite of data-driven techniques they are useful for dealing with non-parametric data when there is insufficient data to use a more explicit technique such as decision trees. The sigmoid and hyperbolic tangent transfer functions used mean that neural nets are rather better at dealing with fuzzy data than the crisp logic of decision trees. Two types of approach are of particular interest in this context. One can be roughly typed as an unsupervised approach, the other as a supervised approached. In some network configurations these can be combined.

The unsupervised approach is exemplified by the Kohonen network or by Self Organising Maps (SOMs) (Kohonen, 1984). A Kohonen network is a single layer of neurodes. Their initial values are set randomly. As each input (training) vector is fed to the layer the neurode with a value closest to the input vector fires. This 'win' by the successful neurode is 'rewarded' by the neurode being allowed to migrate its value closer to that of the input value. Its neighbours are similarly rewarded by being allowed to migrate their values towards the input value, but by a smaller amount. This procedure continues until the Kohonen layer has developed a pattern where similar values are closely adjacent in the layer. This behaviour is similar to that of a decision tree with the neurodes at the end of training being roughly equivalent to the terminal nodes of a tree. procedure is organising the layer in response to similarities in the input vectors. Like decision trees it can result in a number of neurodes, often widely separated, being used to produce a single class in the final thematic map. The problem with this is that no information on the level of discrimination required is being supplied to the training procedure. This is where understanding the link between the problem and the data is very important. The algorithm is grouping the input vectors and has no information on how this relates to a useful output. In some projects this is not a problem. However, if one is trying to produce a thematic map with classes representing the sea, non-forest areas and , say, ten forest types there is a level of

imbalance in the level of discrimination being sought. In order to produce the ten forest types, one would have to produce perhaps as many grassland and non-forest land cover types, probably many more, and as many shallow/deep water classes. This makes the number of neurodes required in the Kohonen layer quite large and consequently, the learning time considerably longer. If this is not done, then one can suppress variability which is needed for subtle discrimination between closely allied classes.

Supervised procedures can avoid this and the commonly used Back Propagation Network is a good model to discuss in this context (Rumelhart & McClelland, 1986). The input vectors are passed down through a multi-layered network. In the training phase, the output layer is compared to the known (or desired) output value or class associated with the input vector. If the output is in error the network weights are altered slightly to reduce the chance of this path being followed next time. If you were a Skinnerian dealing with rats, this could be described as punishing the network for its mistake. Samples are randomly drawn from the training data for as many iterations as are necessary. After a while, the network error rate will tend to stabilise and training can cease. This ability to use the training sample for as many iterations as are necessary is one of the most attractive features of neural nets.

Neural nets of the type discussed here (BPN) work best with a representative learning sample which is made up of vectors which are modal to the desired output classes. If this is done the learning sample size can be kept small. This keeps the degrees of freedom low and increases the level of confidence in the final result.

3.3: Pushing Things to the Limit

Unlike decision trees, BPN can be remarkably tolerant of noisy data if handled carefully. If much of what has gone before sounds like an impossible string of motherhood statements about how we need to clean up our data for these systems, then it is heartening to have a technique which, if used carefully, can cope quite nicely with the realities of data. Indeed, one can even structure investigations which take advantage of this characteristic and are probably not achievable using any other method.

This might best be illustrated using the example of an exercise we carried out across the Liverpool Plains in the Murray Darling Basin. They form part of a highly productive agricultural area, increasingly affected by dryland salinity, which is estimated to cost $10 million per annum in lost agricultural production. Cropping in the area is highly variable, temporally and spatially, as a result of opportunity, summer and winter cropping cycles, and strip and broadacre paddocks. The Liverpool Plains cover an area of 1.2 million ha.

Hydrologically, the Plains are considered as an evaporative basin with a small leak, rather than a fluvial system. Groundwater movement through the basin is complex, and dominated by salinity gradients, microtopographic features and subtle lithological heterogeneities rather than topographic slope. Accurate modelling of this movement would require detailed, and expensive, sub-surface data. An alternative was to attempt to identify empirical evidence of the groundwater movement on the surface and infer its behaviour from that. A first step in doing this was to try to use remotely sensed data. The Liverpool Plains region have a dominant pattern of intensive agriculture. Slight variation in cropping responses due, in the main, to the geochemistry of the soils, is detectable in some places. This is a classic signal detection problem. We are looking for a change in signal on which we can base management strategies. The dominant pattern/signal does not relate to salinity and tends to overwhelm the pattern/signal which may do. In order to provide a more useful management tool we set out to teach a neural network to discriminate the dominant spatial pattern of agriculture, using GIS, and to process the remotely sensed data as though there were no field boundaries and only one crop present.

With an optimising technique to work on this data the number of 'hit' cells in the presence data must be greater than the number of 'miss' cells. Conversely, the number of 'miss' cells in the absence data must be greater than the number of 'hit' cells. If these differences are great, then

the network will converge on an optimum solution without a great deal of trouble. However, the less significant the differences the more care must be taken in setting the learning rate to achieve some sort of convergence.

In order to achieve this we classified SPOT imagery over the area and constructed a polygon coverage of field pattern. Using a modal filter we then labelled these polygons with the modal spectral class within the polygon. This created a simplified image of the land cover, one which 'tends' to be true. There is no assumption that these classes correspond to any particular crop or land cover. We then selected a class which was well represented and was adjacent, at some location or other, to most of the other classes to be the reference class. Using the questionable principle that soil characteristics will not change dramatically over short distances, we then labelled points in each field class as being equivalent to the reflectance value of a neighbouring point in the adjacent reference class field. Because of the necessity to avoid mixels along the field boundaries these two locations were spaced about four cells apart. We then trained a network to learn that the correct reflectance for these points tended to be that of their neighbours across the fence. If this had been true, then all that would have been necessary to do would have been to construct a simple look-up table. Because it only 'tended' to be true, we needed to structure a network learning exercise as though we were dealing with a very poor, or noisy, learning sample. This involved setting a very low learning rate, over a large number of iterations.

The network extracted patterns which appear to represent real geomorphic features, We are now carrying out chemical tests on soil samples to identify the characteristics which are identifiable by the network. This is necessary because the Liverpool Plains are covered by one of the most visually monotonous and homogeneous surfaces it has been my misfortune to deal with. If results from such tests are promising, the network can be further developed and field tested over a larger area. The advantages of this particular methodology, if proven to be a successful predictive tool that can be replicated on scenes from different dates, are that it requires limited input and is independent of vegetation and therefore of growing conditions and cropping cycle, year and stage in season.

Pe[illegible]s its naughty to use the tolerance of the algorithm to [illegible] data in this way, but it does illustrate that a good understanding of the interactions between the algorithm and the data can pay off in unexpected ways.

4: Conclusion

In such a sweeping review as this its difficult to point to a single, tight conclusion. It is however possible to say that times and techniques are changing rapidly and that it is very important not to be distracted from the necessary houskeeping tasks of data management by the fascinating range of new techniques becoming available to us. Indeed, given the characteristics of many of these new techniques in geocomputation, these are perhaps more important than ever.

Bibliography

Aspinall, R. & Lees, B.G. 1995.'Sampling and analysis of spatial environmental data.' in Waugh, T.C. & Healey, R.G. (eds) Advances in GIS Research, Taylor and Francis, Southhampton, 1086-1099.

Austin, M.P. & Smith, T.M. 1989.'A new model for the continuum concept.' Vegetatio, 83: 35-47.

Fitzgerald, R.W. & Lees, B.G., 1993. Assessing the classification accuracy of multisource remote sensing data. REMOTE SENSING OF THE ENVIRONMENT, 41: 1-25.

Fitzgerald R.W., & Lees, B.G. 1994.'Spatial context and scale relationships in raster data for thematic mapping in natural systems.' in Waugh, T.C. & Healey, R.G. (eds) Advances in GIS Research, Taylor and Francis, Southhampton, 462-475.

Kohonen, T., 1984. Self-Organisation and Associative Memory. Springer Verlag, New York.

Lees, B.G. 1994.'Decision trees, artificial Neural Networks and Genetic Algorithms for classification of remotely sensed and ancillary data.' Proceedings 7th Australa-

sian Remote Sensing Conference, Remote Sensing and Photogrammetry Association, Australia Ltd, Floreat, W.A. v1: 51-60.

Lees, B.G. & Ritman, K. 1991. A decision tree and rule induction approach to the integration of remotely sensed and GIS data in the mapping of vegetation in disturbed or hilly environments. ENVIRONMENTAL MANAGEMENT, 15, 823-831.

Moore, D.M., Lees, B.G. & Davey, S. 1991 'A new method for predicting vegetation distributions using Decision Tree Analyses in a Geographic Information System.' ENVIRONMENTAL MANAGEMENT, 15, 59-71.

Rumelhart, D. & McClelland, J., 1986. Parallel Distributed Processing, Vols 1 & 2. MIT Press.

Individuals' cultural code and residential self-organization in the city space

Juval Portugali and Itzhak Benenson
Department of Geography,
University Tel-Aviv, Tel-Aviv,
Ramat-Aviv, 69978, Israel,
phone +972-3-6409178, fax: +972-3-6414148,
e-mail: bennya@ccsg.tau.ac.il

Presented at the second annual conference of GeoComputation '97 & SIRC '97, University of Otago, New Zealand, 26-29 August 1997

1. Introduction.

We consider the city as a complex and open system, that exhibits phenomena of self-organization. We further suggest, that as such a system the city has a special characteristic: its elementary components are human individuals which, unlike the elementary units of non-living and most of the living systems, are themselves self-organizing complex systems. Based on this approach, we have developed a series of agent-based models of city residential dynamics - **City** model, with which we were able to show the emergence of different forms of cultural and economic segregation, and, most importantly, the emergence of a new socio-cultural group in the city space (Portugali, Benenson, Omer, 1994, 1997, Benenson, Portugali, 1995, Portugali, Benenson, 1994, 1995, 1997). Our previous studies were based on the presentation of the individual agent's properties, namely *economic status* and *cultural identity*, as one-dimensional quantitative variables. In this paper, we call off this oversimplifying suggestion, regarding agent cultural identity and consider the latter as a *multidimensional* and *qualitative* variable. Such a representation implies that each individual agent in the model has its own personal "cultural code" (reminiscence in its nature a genetic code), and that the cultural groups of the city consist of individuals with identical cultural code. This formulation allows us to study the *recurrent process of socio-cultural emergence and elimination* in the city.

2. The model.

The model we present elaborates on our previous **City** models. Like them, it consists of two interacting layers - *an infrastructure* submodel, which is an extension of cellular automata and represents the dynamics of the city's physical structure, and a submodel of *free human agents*, which describes the migratory movements of individuals. It differs from past formulation in its definition of the cultural identity of the agents - this is the novel feature we study in this paper.

2.1. The infrastructure submodel.

The infrastructure of **City** is a square **M*M** lattice of cells which symbolizes houses. Each house $\mathbf{H_{ij}}$ can be either occupied by an individual agent or remain empty. We consider a **5*5** square with $\mathbf{H_{ij}}$ in the center as the *neighborhood* $\mathbf{U(H_{ij})}$ of house $\mathbf{H_{ij}}$. Houses differ in their *value* $\mathbf{V_{ij}}$. Each time-step the value of the house is determined anew. When an agent **A** occupies house $\mathbf{H_{ij}}$, its value $\mathbf{V_{ij}}$ is updated in accordance with **A**'s economic status $\mathbf{S_A}$ (see below) and the average value of the neighboring houses in the following way:

$$V^{t+1}_{ij} = (S^t_A + (N(U(H_{ij})) - 1) \cdot \langle V^t_{ij} \rangle_U)/N(U(H_{ij})) \quad (1),$$

where $\langle V^t_{ij} \rangle_U = {}_{Skl}(V^t_{ki} \mid H_{kl} * U(H_{ij}),\ H_{kl} \# H_{ij}))/(N(U(H_{ij})) - 1)$ is an average of houses' values in $\mathbf{U(H_{ij})}$ besides $\mathbf{H_{ij}}$, and $\mathbf{N(U(H_{ij}))}$ is a number of houses in $\mathbf{U(H_{ij})}$.

When a free agent leaves house H_{ij}, and the latter remains unoccupied, the house's value V_{ij} is decreasing at a constant rate:

$$V^{t+1} = d \cdot V^t \qquad (2),$$

where $d < 1$. Here and below we omit , when possible, indices of location.

2.2. The submodel of free human agents.

The individual free human agents of **City** have the ability to estimate the state of the city on its two layers and to behave in line with information regarding **individual**, **local** (referring to the characteristics of the neighborhood's and the neighbors' state) and **global** (referring to the state of the whole city) levels of organization in the city. They immigrate into the city, occupy and change residential locations there, and leave the city when the conditions are unsatisfactory. The agents are characterized by two sets of variables, with which we try to reflect the economic and the cultural characteristics of the human individuals in the city.

2.2.1. Economic characteristics.

The economic state of agent **A** occupying house **H** is given by **A**'s *economic status* S_A. The dynamics of individual's status is described in a simple logistic way:

$$S^{t+1}_A = (R_A \cdot S^t_A \cdot (1 - S^t_A) - m \cdot V^t_H) / \langle V^t \rangle_{city} \qquad (3),$$

where R_A is an *individual rate of economic growth*, that does not depend on **t**, and $m \cdot V^t_H$ is a "mortgage payment", proportional to a house's value.

The local economic information available to individual agent **A**, occupying a house H_{ij}, is given by the economic status of the neighbors and the houses' values in the neighborhood. Formally, the decision of the model individual depends on the difference SD_A between **A**'s status and the mean of the neighbors' status and the unoccupied neighboring houses' values

$$SD^t_A = Abs(S^t_A - P^t_{ij}) \qquad (4),$$

where $P^t_{ij} = ({}_{B}\{S^t_B \mid B \text{ occupies } H_{kl} \text{ e } U(H_{ij}), H_{kl} \# H_{ij}\}$ + ${}_{kl}\{V^t_{kl} \mid H_{kl} \text{ e } U(H_{ij}), H_{kl} \text{ unoccupied}, H_{kl} \# H_{ij}\})$ / $(N(U(H_{ij})) - 1) \qquad (5).$

Below we name SD_A *a local economic tension* of individual **A** at location H_{ij}.

The global economic information available to each individual agent is given by an average of houses' values V^t_{ij} over the city:

$$\langle V^t \rangle_{city} = {}_{kl}\{V^t_{kl} \mid k, l \text{ e } [1, M]\} / (M{*}M) \qquad (6).$$

2.2.2. The cultural code.

Each human individual enters the world with an inherited genetic code, which pre-program his/her possibilities to behave and interact with other individuals when creating groups or societies. Inspired by this perspective, we suggest that every individual agent in our model enters the city with a "cultural code", which defines its possibilities for residential behavior and interactions with other agents. In genetics of qualitative features as well in studies of artificial life, it is common to represent the individual's genotype by means of a high-dimensional binary vector (Banzhaf, 1994). Below, we introduce the cultural code of an individual agent in the same manner. As emphasized in our previous papers, and at the outset of the present one, we suggest that human agents are characterized by their ability to vary, and, consequently, self-organize, in line with the dynamics and evolution of the system they belong to. We, therefore, suggest, that the cultural code of an agent and its residential behavior can change through its interaction with its neighbors, neighborhood, and the city as a whole.

2.2.3. Cultural characteristics.

The cultural code of an individual **A** is described by the **K**-dimensional Boolean vector $C_A = (c_{A,1}, c_{A,2}, c_{A,3}, \ldots, c_{A,K})$, where $c_{A,k}$ e $\{0, 1\}$, $k = 1, 2, 3, \ldots, K$. As a result, individuals of 2^K different cultural identities might exist in the city. Individuals **A** and **B** have different identities when vectors C_A and C_B differ in at least one component. Quantitatively we measure this by difference **r** between **A**'s and **B**'s identities:

$$r(C_A, C_B) = {}_{k}(c_{A,k} \text{ XOR } c_{B,k}) / K \qquad (7).$$

The representation of *local cultural information* is related to the notion of *local spatial cognitive dissonance* of free agent **A**. Applying the general definition (Portugali, Benenson, 1995, Haken, Portugali, 1995) to the multidimensional presentation of cultural identity we define local spatial cognitive dissonance CD_A of agent **A**, occupying house H_{ij}, as an average of the differences between **A**'s identity and the identities of his neighbors:

$$CD^t_A = {}_{sB}\{r(C^t_A, C^t_B) \mid B \text{ occupies } H_{kl} \in U(H_{ij}), H_{kl} \# H_{ij}\}/(N^t_{oc}(U(H_{ij})) - 1) \quad (8),$$

where $N^t_{oc}(U(H_{ij}))$ is the number of occupied houses in $U(H_{ij})$.

If individuals similar to **A** in their cultural codes are segregated in the city at a certain degree, then their spatial distribution might affect the behavior of **A**. For this purpose we define *a global cultural information* GD_A, available to free agent **A**, about the level of residential segregation of the individual agents of identity C_A. We use the Lieberson (1981) segregation index $LS_{X,Y}$ to characterize the level of segregation of a certain group **X** relative the other group **Y**. $LS_{X,Y}$ is a probability for individual **A** that belongs to group **X** and located at house **H**, to meet a member of group **Y** within **U(H)**. The complete information on the residential segregation in the ***City*** at iteration **t** is given by the $2^K{*}2^K$ matrix of Lieberson segregation indices $LS^t_{X,Y}$ for each pair of cultural identities **(X,Y)**. To decrease the enormous dimensions of this description we suggest below that agent **A**'s behavior depends on the global level of segregation of its cultural group relative all the other individuals *taken together*, and denote the corresponding value of Lieberson index as LS^t_A. The dimension of the latter description equals to the number of identities, i.e. 2^K. The values of LS_A below **0.2** corresponds to visually random distribution of agents of identity C_A, while the values above **0.8** correspond to one or several domains occupied by the these individuals almost exclusively. Quantitatively, we describe the global cultural information an agent **A** accounts for as:

$$GD^t_A = \max\{0, (LS^t_A - LS^*)\}/(1 - LS^*) \quad (9).$$

Here LS^* is the value of Lieberson index that corresponds to visually segregated pattern, and below we set LS^* equal to **0.4**.

We suppose, that local and global information influence agent's cultural identity in alternative ways. High local cognitive dissonance CD^t_A forces an individual agent **A to change** its *cultural* identity, and an **A**'s *reaction to the* local cognitive dissonance is characterized in the model by *a sensitivity* $L_A \in [0, 1]$. In the opposite direction, high level of segregation of individual agents of identity C_A, forces **A to preserve** its current identity, and an agent's reaction to the global segregation is characterized by *a sensitivity* $G_A \in [0, 1]$. Below we suggest that L_A and G_A are inherent properties of **A** and do not depend on **t**.

The change in an agent's cultural identity thus depends on two controversial tendencies. The cultural identity of an agent **A** can be changed when the local tendency to change an identity exceeds the global tendency to preserve it, i.e. when $L_A \cdot CD^t_A > G_A \cdot GD^t_A$. If the latter is true, then the probability that the **i**-th component of C_A will be changed is proportional to the absolute value of the difference between the fraction of this component among **A**'s neighbors and its value for **A**. Additionally, we introduce the possibility for a "cultural mutation" with probability r_m per component of identity. As a result, for an agent **A** of identity $C_A = (c_{A,1}, c_{A,2}, \dots, c_{A,i}, \dots, c_{A,K})$, occupying house H_{ij}, the probability of change in the **i**-th component $c_{A,i}$ of C_A to its negation, i.e. from unit to zero or vice versa, is

$$p_{A,i} = \max\{0, (L_A \cdot CD^t_A - G_A \cdot GD^t_A) \cdot ((Abs(f_i - c^t_{A,i}) + r_m)/(S_k Abs(f_k - c^t_{A,k}) + r_m \cdot K)\} \quad (10),$$

where f_k is a frequency of **not** $c^t_{A,k}$ in the cultural identities of **A**'s neighbors at iteration **t**:

$$f_k = {}_{sB}\{ c^t_{B,k} \text{ AND } (\text{NOT } c^t_{A,k}) \mid B \text{ occupies } H_{kl} \in U(H_{ij}), H_{kl} \# H_{ij}\}/(N^t_{oc}(U(H_{ij})) - 1) \quad (11).$$

We suppose that only one component of cultural identity can be changed at a time-step.

2.3. Model dynamics: trade off between migration and individual's change.

According to the flow-chart (Fig. 1), at every iteration, each free agent **A** in the city decides whether to move from, or to stay at, its present location. As it is shown in Fig. 2, the *probability to leave* a house *increases monotonously*, and the probability to occupy a new house *decreases monotonously* with an increase in either individual's economic tension SD_A (see formula 4) or cultural dissonance CD_A (see formula 8).

We calculate the probability that agent **A** will leave its house as:

$$p(SD^t_A, CD^t_A) = 1 - (1 - p_e(SD^t_A)) \cdot (1 - p_c(CD^t_A)) \quad (12),$$

and the probability that **A** will occupy a vacant house H_{ij} as:

$$q(SD^t_A, CD^t_A, H_{ij}) = q_e(SD^t_A, H_{ij}) \cdot q_c(CD^t_A, H_{ij}) \quad (13),$$

where **p** denotes the probability to leave a house, **q** denotes the conditional probability to occupy a vacant house H_{ij}, when it is the only possible choice, and indices **e** and **c** denote economic and cultural components. A vacant house H_{ij} is attractive for an agent **A**, when at least four houses in $U(H_{ij})$ are occupied. For details see Portugali, Benenson, Omer (1997).

The conjunction between individual, local and global factors, can lead individual agent **A** to decide to continue to occupy house **H** in spite of high economic tension and cultural dissonance. The reason, for example, might be a lack of attractive vacant houses in the city. The basic suggestion of the **City** model is that in such a situation the dissonance is resolved either by leaving the city, or by *change in the properties of the free agent itself*.

2.3.1. Free agent's behavior under increasing economic tension.

The change in the individual's status is an inherent source of the **City** economic dynamics. If an agent's status changes significantly faster, or slower, than the average status of the neighborhood, the agent either tries to migrate to another location or "goes bankrupt" according to (1) and migrates out of the city.

2.3.2. Free agent's behavior under *increasing cultural cognitive dissonance.*

An inherent source of the **City** cultural dynamics is a mu-

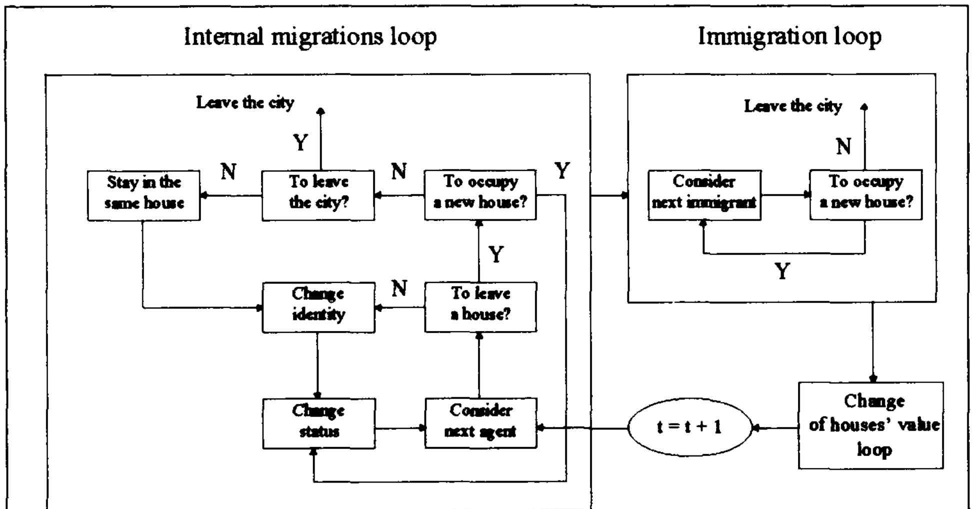


Figure 1

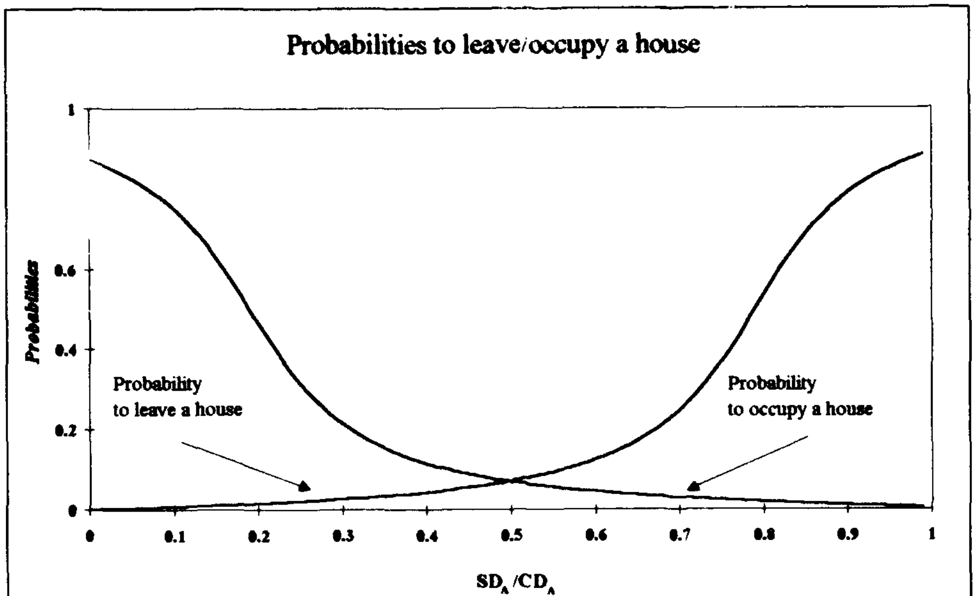


Figure 2

tation process, that prevents the *City* fro[illegible]ng culturally homogeneous. An individual [illegible] in an heterogeneous neighborhood of non-zero [illegible]ance, either succeeds to change residence, or fails and, thus, either changes an identity towards the "modal" identity of the neighbors (Fig. 1), or preserves its current identity due to high level of segregation of agents of similar identity in the city (see formula 10). Unlike the changes in the one-dimensional economic status, the changes of agents' cultural identity *do not decrease* the cultural diversity of the city when $K > 1$. As an example consider the agents located at a boundary between two segregated groups of individuals $(0, 0, 0, \ldots, 0)$ and $(1, 1, 1, \ldots, 1)$. According to (10), there is a high probability that the identity of, say, the $(0, 0, 0, \ldots, 0)$-agent will change to a new one with a unit at one of the components and, thus, will differ from identities of the agents of *both* groups. This salient consequence of multidimensional representation of C_A determines most of the results below.

2.3.3. Emigration.

We have stated above that an individual, whose economic status reaches zero, leaves the *City*. A free agent that failed to reside might (1) leave the City with probability p_U; (2) change cultural identity with the probability given by (10); and (3) stay at his current location and do not change at all.

2.3.4. Immigration.

At every time-step, a constant number of I individuals try to enter the city from outside and to occupy a house in it. The economic status S_i and growth rate R_i of each immigrant I are assigned randomly and independently. The distribution of S_i is a normal truncated on $[\min_{city}(S^{t-1}), \max_{city}(S^{t-1})]$ with a mean equals to the instantaneous mean status of the city agents $<S^{t-1}>_{city}$ and constant CV. The distribution of R_i is a normal truncated on $[0, R_{max}]$ and does not depend on t.

Cultural identity of the immigrants is assigned at random, in proportion to the current fractions of agents of each of the 2^K possible identities.

3. *Results*.

The aim of our model is to examine the process of socio-cultural emergence in the city, the inhabitants of which can vary in their cultural identity according to potentially infinite number of traits. To qualify as a newly emerging socio-cultural entity, a group of individuals must fulfill simultaneously three conditions (Portugali, Benenson, Omer, 1997). At the **individual level** the members of the group must have the same cultural identity, at the **local level** most of the group members should be located within neighborhoods of their own, and at the **global level** the number of group members and their spatial segregation have to be sufficiently high.

Our previous studies (Portugali, Benenson, Omer, 1994, 1997, Benenson, Portugali, 1995, Portugali, Benenson, 1994, 1995, 1997) show that different sets of parameters might generate three kinds of residential dynamics in the ***City***. One is a "random" city, another is a "homogeneous" city, in which most of the agents belong to the same group, and the third is characterized by a complex structure. All these regimes are observed in the present study too, and, below, we deal with the set of parameters that entails the most interesting "structured" dynamics. In this paper we are specifically interested in the question of whether the residential distribution of the individual agents in the city evolves towards a state that can be called "persistent" in some respect and, if so, what are the characteristics of this state. In particular, what is the number, and the level of segregation, of the emerging socio-cultural entities; are they fixed? do they vanish in time? what is their "life-history"? Below we concentrate on cultural identity only and, therefore, set $p_e(SD^c_A) = 0$ and $q_e(SD^c_A, H_t) = 1$.

3.1. Parameters' value and initial conditions.

The scenarios we run share the following conditions:

1. ***City*** is a **40*40** lattice.
2. Initially, at $t = 0$, each cell within a circle of 3-cell diameter, located at the center of city lattice, is randomly occupied by individuals of all-zero cultural identity **(0, 0, 0, ..., 0)**.
3. Immigration rate **I** equals **4**, or **0.25%** of maximum number of the city residents - **1600**.
4. Probability p_U to leave the city, when failing to occupy a new house, equals **0.075**
5. Distributions of sensitivities **L** and **G** are uniform on **[0, 1]**. They are assigned to the agents independently of each other.
6. Mutation rate r_m is **0.02**.
7. Threshold group size sufficient to recognize a group as an "entity" is **40** individuals (enabling up to 40 different identities to exist simultaneously in the city).

At present, our computer allows us to study the system behavior when the dimension of the cultural identity vector C_A is less or equal to **5**. The question of whether the case of $K = 5$ is representative of a higher-dimensional C_A, will be studied further.

The results below are common for five repetitions of each scenario.

3.2. Presentation of the *City* patterns.

There exist certain difficulties in presenting the spatial characteristics of the city when a cultural identity is a multidimensional vector. To present the image of the city, we use below three kinds of maps. The first one is a distribution of agents' cultural identity, with each identity marked by its own color. This presentation is the most detailed one, but is unacceptable for $K > 2$, in view of high number and non-linear ordering of identities. The second type of maps is that of difference $r(C_A, C_0)$ between the identity C_A of agent **A**, occupying house **H** and an *a priori* chosen identity that equals, say, $C_0 = (0, 0, 0, \ldots, 0)$. This map shows the effects that do not depend on **K**, but its disadvantage is that several different identities C_A can equally differ from the selected for comparison. The third map is that of a distribution of cultural cognitive dissonance of the residents. This map is a surrogate of Stability-Instability Surface (Portugali, Benenson, Omer, 1995) in the sense that the higher is the dissonance, the higher is the chance that the state of a given house will change.

Before proceeding to the analysis, let us point out that the dynamics of the distribution of cultural identity depends

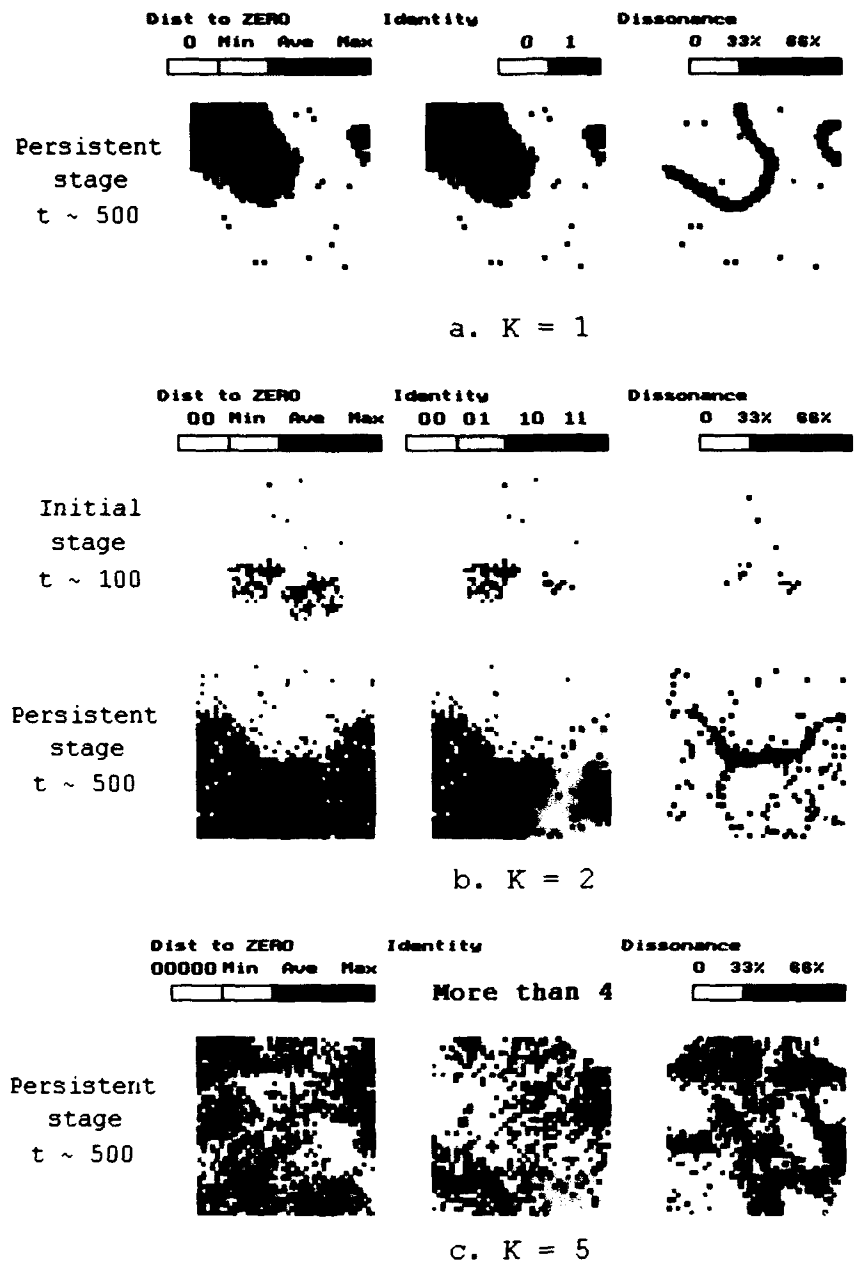

Dist to ZERO
0 Min Ave Max
Identity
0 1
Dissonance
0 33% 66%
Persistent stage t ~ 500
a. K = 1
Dist to ZERO
00 Min Ave Max
Identity
00 01 10 11
Dissonance
0 33% 66%
Initial stage t ~ 100
Persistent stage t ~ 500
b. K = 2
Dist to ZERO
00000 Min Ave Max
Identity
More than 4
Dissonance
0 33% 66%
Persistent stage t ~ 500
c. K = 5

Figure 3

on the number K of its components. In general, an increase in K increases the "resolution" of identity, but keeps the range of its variability. We mean here that according to (7) the maximal possible value of $r(C_A, C_B)$, i.e. difference between individual A of identity C_A and individual B of the opposite identity C_B ($C_A = (0, 0, 0, \ldots, 0)$ and $C_B = (1, 1, 1, \ldots, 1)$, for instance) remains equal to unit, no matter what K is.

3.3 Model dynamics for low-dimensional cultural identity: K = 1 and K = 2.

The case of K = 1 corresponds to our previous analysis of residential segregation between two cultural groups (Portugali, Benenson, Omer, 1994). The city dynamics in that case entailed a fast self-organization of (0)- and (1)-identities within two or several segregated patches. The boundaries between the homogeneous patches remain the areas of instability, with intensive exchange of individuals (Fig. 3a, compare to Portugali, Benenson, Omer, 1994).

When K equals two, the dynamics of the city still resemble some of our previous results (Portugali, Benenson, 1997, Portugali Benenson, Omer, 1997). At the beginning of the runs, in line with the restriction of mutation process by one component per iteration, only (0, 1)- and (1, 0)-agents emerge. The numbers and the level of segregation of the initial (0, 0)- and of new (0, 1)- and (1, 0)-identities reach the levels satisfying the conditions of socio-cultural emergence, to t ~ 100, when the fraction of unoccupied locations in the city is at a level of 25%. The agents of (0, 1)- or (1, 0)-*identities that change it to a* (1, 1) because of mutation or dissonance with the neighbors, still have the vacant houses to reside. As a result, the (1, 1) socio-cultural entity emerges in the City (Fig. 3b) in all of the model runs to t ~ 400. In parallel, the number of vacant houses tends to approach zero, and strong competition for houses turns to be the factor that defines the survival of the entities. In general, the survival of a certain entity is defined by the position and the size of the domains, it occupies. The high value of the perimeter/area relation, as well as the common boundary with an opposite entity (e.g. (0, 1) for (1, 0)-agents) decreases the chance that the entity will persist. As a result, in a long run (we stopped the simulations at t = 2500) the number of socio-cultural entities, existing simultaneously in the city for K = 2, fluctuates between three and four, and the life-span of the entities is of the order of 500 iterations.

Let us now skip an intermediate cases of K equals 3 and 4, and proceed with K = 5.

3.4. Model dynamics for high-dimensional cultural identity: K = 5.

3.4.1 Initial stage of the model dynamics.

The number of possible identities for this case is $2^5 = 32$. The first mutant agents belong to one of five "close-to-zero" identities, which are characterized by unit at one of the components and zeros at the rest of them and, compared to K = 2, it is *not necessary* that all of them will emerge at the first stage of the city development. In the five runs we did, their number vary between two and four.

3.4.2. Persistent dynamics of the city.

The entities that emerge first determine the further dynamics of the city. In a way similar to the case of K = 2, the boundaries between two homogeneous domains (occupied by the entities that emerged at the first stage) and the *heterogeneous domains*, occupied by the agents of varying identities, are areas of instability. The agents located there, either leave their houses or change their identity. None of the properties of the certain socio-cultural entity currently existing in the City can be predicted in a long run. As a result, we cannot follow the qualitative fate of certain identity, but still are able to understand and predict are the properties of the model city as a complex self-organizing system:

1. The persistent city structure is characterized by a mixture of spatially homogeneous domains, the population of which forms socio-cultural entities, and domains that are heterogeneous at different level. The former cover about half of the city for K = 5 (Fig. 3c).

2. A limited number of cultural entities can exist in the city simultaneously (Fig. 3c, Fig. 4).

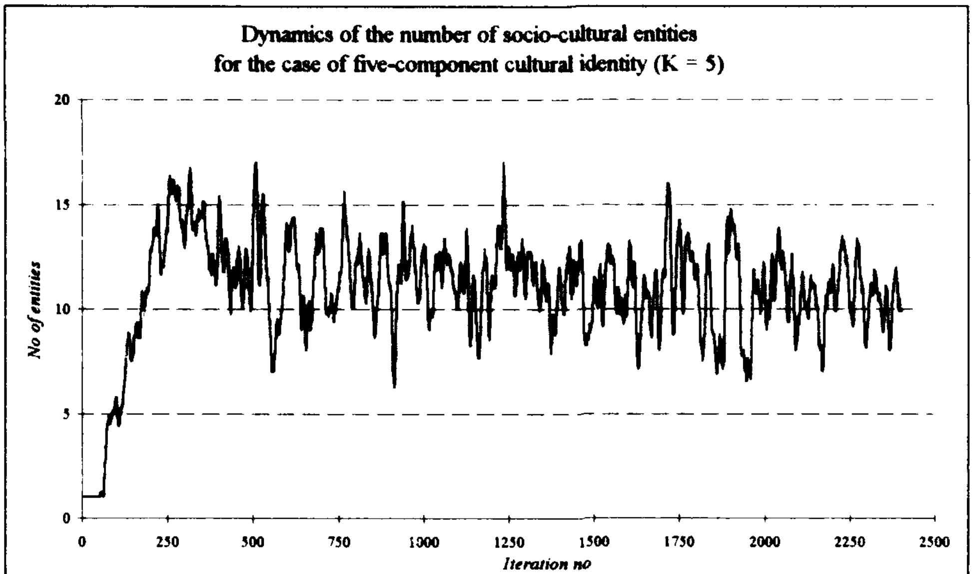


Figure 4

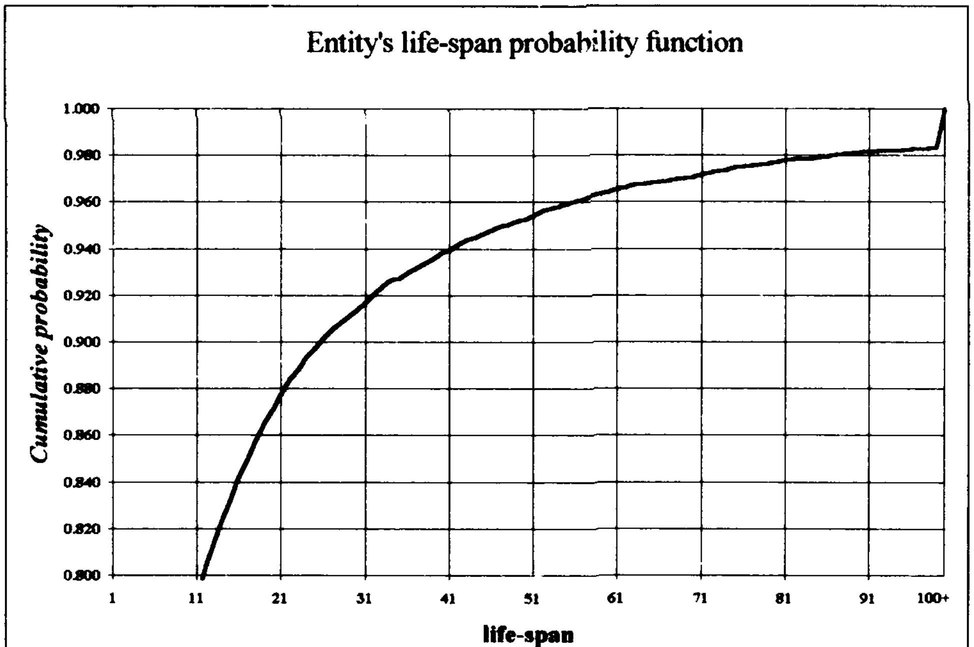


Figure 5

3. The life-span of socio-cultural entity is finite and the entities replace each other in the city space. About 20% of the entities persist in the city 11 iterations or longer and 10% persist 25 iterations or longer (Fig. 5).

4. The distribution of cultural differences $r(C_A,C_0)$ between the cultural identity C_A of agent **A**, and certain "basic" cultural identity C_0 (C_0 = (0, 0, 0, 0, 0) for the maps we present here) is self-organizing as well (Fig. 3). This distribution has two opposing characteristics. First, in general, the difference $r(C_A,C_0)$ increases with the increase in the distance from the location of the agents of the C_0 identity. Second, *non-linear ordering of the identities* implies the emergence of the adjacent areas of entities C_A and C_B, that equally differ from C_0 (i.e. $r(C_A,C_0) \sim r(C_B,C_0)$), but differ also among themselves (i.e. $r(C_A,C_B)$ is high). See, for instance, the bottom part of Fig. 3c, where the boundary between yellow and violet domains is an area of high dissonance. This property, determined by the multidimensional and quantitative nature of cultural identity of the model agent, limits the *City's* instability *from below*. With an increase in the segregation in the city, its instability does not converge to zero (Fig. 6) and several unstable zones are preserved. We can say, thus, that the city is *self-organizing* and evolving toward *critical* internal structure, that preserves the ability to changes.

4. Conclusions and Discussion.

Our research is based on the idea, that an individual human agent is able to change him/herself, depending on information at different levels of self-organizing city structure. Such an idea implies the possibility of socio-cultural emergence in the city (Portugali, Benenson, 1997, Portugali, Benenson, Omer, 1997). In this paper, we introduce the notion of "cultural code" which describes the individual as a multidimensional and qualitative unit. From this perspective, follows three new qualitative phenomena. First, *recurrent self-organization, emergence and extinction* of the socio-cultural groups in the city. Second, *only a limited number of cultural entities* (from a large number of possible ones) can *exist simultaneously* in the city space. Third, the city as a

City's instability, given by fraction of agents that want to leave a house, vs number of socio-cultural entities

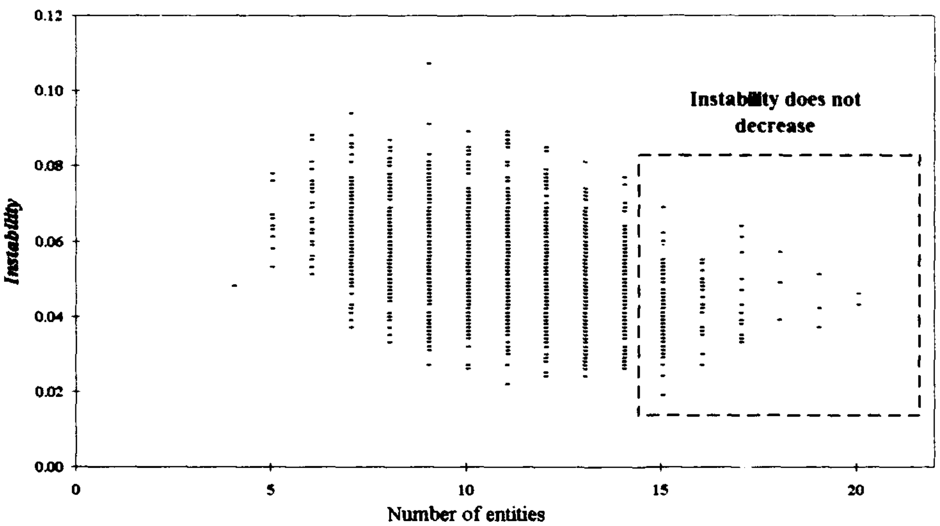


Figure 6

whole tends towards a self-organized critical state that preserves cultural instability. As a result, the city cultural landscape is a mixture of a few homogeneous domains, each one occupied by individuals of a certain socio-cultural identity, and areas of heterogeneous population. The identities of the existing socio-cultural entities and their further evolution depend on the emerging situation and cannot be predicted in advance.

An important question we do not discuss in the present paper concerns the consequences of the interrelations between self-organizing cultural and economic city structures. The evolution of the latter has intensively been studied during the last three decades, when most of the recent efforts are performed within the framework of the Cellular Automata models. The CA modeling clearly demonstrates effects of self-organization in the city space. (Batty, Xie, 1994, Itami, 1994, Benati, 1997, Sanders et al, 1977). The resolution of CA models is at the level of several parcels of land, and the their outcome is in good agreement with the dynamics of the real cities (Wu, 1996, White, Engelen, Uljee, 1997). The number of the cell states in CA models, which usually refer to land uses (housing, industry, commerce, etc.), is always predetermined with the implication that no new form of land use can emerge in the city. Our agent-based models operate at the level of separate individuals and houses and enable the possibility of emergence of a qualitatively new groups in the city space (Benenson, Portugali, 1995). When cultural and economic characteristics of agents are considered together, we can demonstrate coherent self-organization of the city economic and cultural landscapes (Portugali, Benenson, 1995). The phenomenon of socio-cultural emergence provides a low-resolution mechanism that enables qualitative bifurcations of the city spatial dynamics (Haken, Portugali, 1995). The construction of a comprehensive model, that combines cellular automata with the agent-based approach can be a further step towards understanding the dynamics of the city as a self-organizing system.

5. Bibliography.

1. Banzhaf W, 1994, "Self-organization in a system of binary strings", In *Artificial life IV*, Eds. R.A.Brooks and P. Maes (MIT Press, Cambridge, MA), pp. 109-118.

2. Batty M, Xie Y, 1994, "From cells to cities" *Environment and Planning B*, v21, pp. 531-548.

3. Benati S, 1997, "A cellular automata for simulation of competitive location" *Environment and Planning B*, v24, pp 205-218.

4. Benenson I, Portugali J, 1995, "Internal vs. external spatial information and cultural emergence in a self-organizing city", In *Spatial information theory: a theoretical basis for GIS*, Eds. A.U. Frank, W. Kuhn (Springer-Verlag, Berlin), pp 431-441.

5. Haken H, Portugali J, 1995, "A synergetic approach to the self-organization of cities and settlements" *Environment and Planning B*, v22, pp 35-46.

6. Itami R, 1994, "Simulating spatial dynamics: cellular automata theory" *Landscape and Urban Planning*, v30, pp. 27-47.

7. Lieberson S, 1981, "An asymmetrical approach to segregation", in *Ethnic segregation in the cities*, Eds. C. Peach, V Robinson, S. Smith (Croom Helm, London), pp 61-82.

8. Portugali J, 1996, "Inter-representation networks and cognitive maps", in *The construction of cognitive maps*, Ed. J. Portugali (Kluwer, Dordrecht), pp 11-43.

9. Portugali J, Benenson I, 1994, "Competing order parameters in a self-organizing city", in *Managing and marketing of urban development and urban life*, Ed. G.O. Braun (Dietrich Reimer, Berlin), pp 669-681.

10. Portugali J, Benenson I, 1995, "Artificial planning experience my means of a heuristic sell- space model: simulating international migration in the urban process" *Environment and Planning A*, v27, pp 1647-1665.

11. Portugali J, Benenson I, 1997, "Human agents between local and global forces in a self- organizing city" In *Self-*

organization of complex structures: from individual to collective dynamics Ed. F. Schweitzer (Gordon and Breach, London). pp 537-546.

12. Portugali J, Benenson I, Omer I, 1994, "Socio-spatial Residential Dynamics: Stability and Instability within a Self-Organizing City" *Geographical Analysis*, v26, pp 321-340.

13. Portugali J, Benenson I, Omer I, 1997, "Spatial cognitive dissonance and socio-spatial emergence in a self-organizing city" *Environment and Planning B*, v24, pp 263-285.

14. Sanders L, Pumain D, Mathian H, Guerin-Pace F, Bura S, 1997, "SIMPOP: a multiagent system for the study of urbanism" *Environment and Planning B*, v24, pp 287-306.

15. White R, Engelen G, Uljee I, 1997, "The use of constrained cellular automata for high-resolution modeling of urban land-use patterns" *Environment and PlanningB*, v24, pp. 323-343.

16. Wu F, 1996, "A linguistic cellular automata simulation approach for sustainable land development in a fast growing region" *Comput., Environ. and Urban Systems*, v20, pp. 367-387.

Modelling hydraulic, sediment transport and slope processes, at a catchment scale, using a cellular automaton approach.

T.J. Coulthard, M.J. Kirkby & M.G. Macklin
School of Geography, University of Leeds
Leeds, LS2 9JT
United Kingdom
Tel: +44(0)113 2333326, Fax: +44(0)113 2333308
E-mail: T.Coulthard@geog.leeds.ac.uk

Presented at the second annual conference of GeoComputation '97 & SIRC '97, University of Otago, New Zealand, 26-29 August 1997

Abstract.

There have been recent advances in the numerical modelling of hydraulic and sediment transport processes at a fine scale, but the ability to extrapolate these advances to a larger scale is rarely realised. Existing approaches have been based upon linked cross sections, giving a quasi 2-d view, which is able to effectively simulate sediment transport for a single river reach. A catchment represents a whole discrete dynamic system within which there are channel, floodplain and slope processes operating over a wide range of space and time scales. A Celluiar Automaton (CA) approach has been used to overcome some of these difficulties, in which the landscape is represented as a series of fixed size cells. At every model iteration, each cell acts only in relation to the influence of its immediate neighbours in accordance with appropriate rules.

The model presented here takes approximations of existing flow and sediment transport equations, and integrates them, together with slope and floodplain approximations, within a cellular automaton framework. This method has been applied to the Catchment of Cam Gill Beck (4.2 km^2) above Starbotton, upper Wharfedale, a tributary of the River Wharfe, North Yorkshire, UK.

This approach provides for the first time a workable model of the whole catchment at a meso scale (1m). Preliminary results show the evolution of bars, braids, terraces and alluvial fans which are similar to those observed in the field, and indicates the emergence of significantly non-linear behaviour.

Introduction

Fluvial sediment transport and the supply of sediment to and from the floodplain are the most important processes in the evolution of a catchment. For this and other reasons, fluvial models, operating at a variety of scales, have taken a precedence in geomorphology. These range from the three dimensional modelling of circulation surrounding a confluence, detailed two dimensional finite element grids of water surface profiles (Nicholas 1997, Bates et al 1997) and the more 'classic' one dimensional approach of calculating over cross sections, such as HEC II. Most appear successful, but due to the complexity of solving the complex Navier-Stokes equations used, are computationally restricted to operating in a confined area. They also fail to account for processes outside of this study reach, such as mass movement, hydrology and changes in upstream sediment supply.

Other authors, Howard (1994, 1996), Polarski (1997) take a different approach, placing the emphasis on the slope processes. Howard simplifies channel operations to a sub grid cell process, with values for width and depth calculated using empirical relationships. This approach allows the aggradation and degradation of the channel, in the con-

text of the whole catchment, but does not allow the formation of terraces, a flood plain stratigraphy, or differing channel forms which geomorphologists use to interpret past environmental change.

Whilst both of these approaches are fruitful, the former, hydraulic approach trades catchment scale realism for local flood plain accuracy, whereas the latter sacrifices channel accuracy for realism at the catchment scale. Two reasons for this split can be identified. Firstly, numerical flow modelling mainly comes from a strongly engineering background, where the prime consideration is the channel. The second reason is scale.

When examining a topic as complex as landscape evolution, there are numerous processes acting over a wide range of time and space scales. These range from the movement of a pebble in a split second, to the creep on a mountainside over thousands of years. The importance of a mass landslide in changing the landscape is obvious, but should we ignore the pebble's movement? If we assume our landscape to be a chaotic system, highly sensitive to initial conditions, then the pebbles' action is important, as is the butterfly effect to a climate modeller. Lane et al (1997) seems to confirm this idea, suggesting that fluvial system behaviour is highly dependant upon its context. This presents a major problem for a modeller in selecting an appropriate level of resolution. For example, if studying the Rhine Basin, how far should we account for the turbulence generated by the movement of a 5mm clast? In principle the answer is not clear, as there are critical moments when it influences the outcome, but in practice computational limits effectively exclude such a high level of detail.

Incorporating small scale processes in a catchment model is troublesome, because of these scale ranges. The computationally intensive nature of finite element methods makes their use impracticable over the long timescale that slope influences require (>1000 years), and it is similarly impossible for them to provide models for the full spectrum of flood events. Furthermore, over the course of a flood, catchments are spatially dynamic. Stream heads may extend, new tributaries and channels may form. For hydraulic modelling this creates numerous problems, as changes in bed/floodplain topography and spatial changes in the network require a frequent re-definition of the mesh of nodes used, which is highly time consuming, especially if a curvilinear approach is used.

In this paper, a cellular automaton (CA) model, simple in concept yet complex in implementation, is applied to an entire small upland catchment. This model aims to reconcile scale issues by dividing the catchment into uniform 1m ^{2} grid cells. This resolution is chosen as being small enough to allow representation of fluvial processes, yet large enough to encompass a whole catchment. Furthermore, to resolve temporal scale problems a variable time step is used which is dependant upon the erosion rates. This allows the representation of small scale processes such as fluvial erosion, yet incorporates the long term effects of vegetation change and soil creep. This model is being developed as part of on-going research to investigate the relative effects of climate change and humans influence on the upland landscape over the Holocene (Coulthard *et al* 1996, 1997, Macklin & Lewin 1993). In this paper the Authors wish to :

1. Focus on the models unique application at this scale.
2. Investigate examples of non linear behaviour in the relationships between processes.
3. To consider an appropriate choice of scale, for models of environmental change.

Method.

The model is applied to the catchment of Cam Gill beck, a tributary of the River Wharfe, above the hamlet of Starbotton, North Yorkshire, UK. The CA method used and details regarding its implementation are described in full by Coulthard *et al* (1996, 1997) but sumarised below.

The catchment was digitised from 1:10 000 scale Ordnance Survey map contours. This data, with additional EDM surveyed detail for the valley floor was combined using the TOPOGRID command in ARC-INFO to create a 1 m^2 resolution DEM, of 4.2 million points (figure 1.). Within this topographic representation, each grid cell has proper-

ties of elevation, discharge, vegetation, water depth, and grainsize. For every model iteration, these values are altered in accordance only to their immediate neighbour and four sets of processes. The first component is a model of hill slope hydrology, using an adaptation of TOPMODEL (Beven & Kirkby 79) with an exponential soil moisture store. The second input is a hydraulic routing scheme, utilising bed slope and calculating depth with an adaptation of Mannings formulae. Thirdly, fluvial erosion and deposition using the Einstein-Brown (1950) equation, applied to five different grainsize fractions incorporated with a 3 strata active layer system similar to that used by Parker (1990) and Hoey & Ferguson (1994). Finally, mass movement rates are calculated, incorporating a factor of safety which changes with the soil saturation.

Two main scenarios have been applied to the model. Firstly fifteen floods of equal magnitude, equivalent to a bankfull discharge have been simulated, to show cumulative changes in sediment discharge and morphology. Secondly, a larger flood approximating to a 5 year flood event was simulated.

Results.

Figure 2 shows the results of running 15 floods of approximately bankfull discharge through the upper part of the catchment. This graph shows two values, firstly the amount moved in each flood and secondly the amount removed from the catchment. The initial conditions were with an 'untouched' catchment where every cell had the same grainsize content. This meant that for the first few runs large amounts of material were removed because the channel was armouring itself from these initial conditions and had a high sediment availability. Subsequent to this peak, the catchment displays a non linear pattern of behaviour, with unrelated peaks in the sediment discharge. This may be attributed to the movement of 'slugs' (Nicholas *et al* 1995) of sediment down stream, and the consequent remobilisation of these, in later floods. These peaks in activity can be also be linked to the input of landslides. Mass movement producing an input of fines into the system. When monitoring the model's operation, the activity in the catchment corresponds to that of the hydrograph. Little happens until the peak of the hydrograph occurs then there is a flurry of activity as sediment is mobilised. This then decreases with the falling limb. There are however episodes of activity during fairly low flow times. This is again attributed to the input of mass movement from the slopes.

Figures 3 to 6 show the confluence section as indicated in Figure 1. These show the confluence of the main two upland channels. Figure 3 shows the 'initial conditions of the area, where a small discharge has been run down the catchment, resulting in the definition and formation of channels. Figure 4 shows the same region after the 16 floods outlined in Figure 2 above. Figure 5 shows again the same area, but after 1 large flood of approximately 5 year return interval. These three views show the activity of several processes. The floods have led to the development of a 'fan' like structure at the base of the right hand tributary, produced by fines from the upland areas. This has caused the widening of the channel opposite and downstream. A multiple channel has formed here, due to the large sediment influx, the channel diverging and converging. Figure 6 corroborates these observations, showing the grainsize distribution for the section after the 16 floods. This shows an 'armouring' down the centre of the multiple channels and a 'glut' of fine material deposited at the base of the fan.

Figures 7 a & b, show two plan-views of one small section of 80 by 30m, as outlined on Figure 1. Flow is from top to bottom. On the right are four cross sections corresponding to the sections on the grainsize chart. This is a lower part of the channel seen after the 16 floods mentioned above. Here, two distinctly different formations have occurred. In the upper two sections, the flow emerges from a narrow constrained section into a wide valley floor. Consequently there has been deposition, with the formation of a coarse deposit on the right side. 30m downstream, where the system is eroding, removing the deposits from above, the opposite has occurred, where there is a fine deposit on the left of the channel. These features are very similar to a 'boulder berm' and side bar / terrace, in both

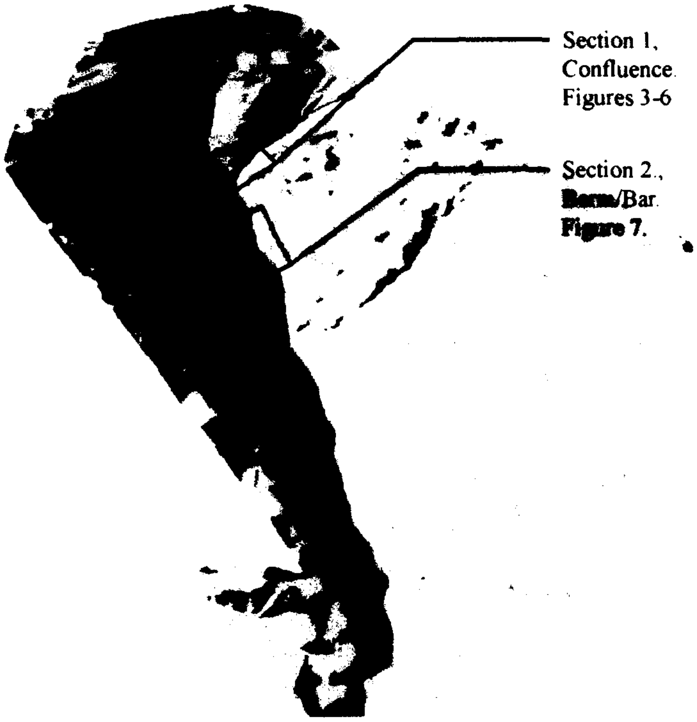


Figure 1. Draped image of Starbotton DEM. Scale 1600 by 2800m.

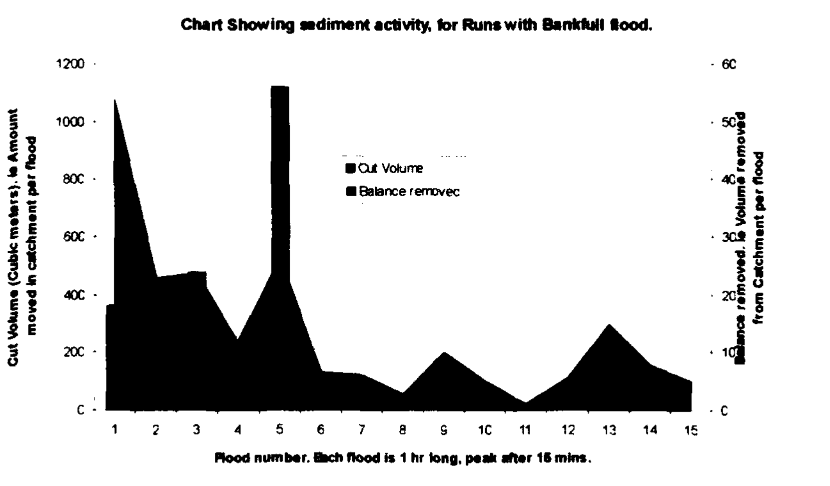


Figure 2. Graph showing volume of sediment moved and removed from the catchment for each flood.

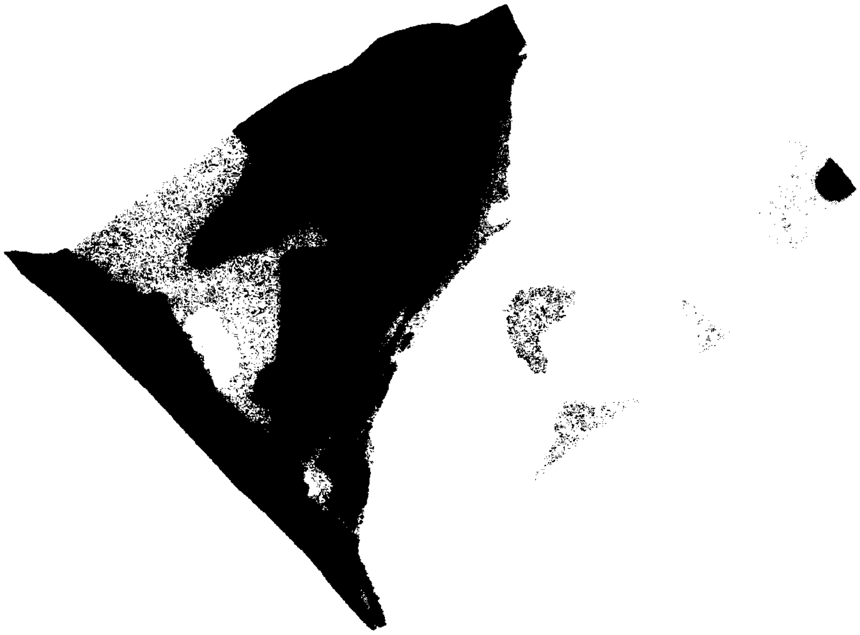


Figure 3. Confluence section before flood series.

Fig 4. After 16 floods of bankfull discharge.

Figure 5. After a '5 year' flood event.

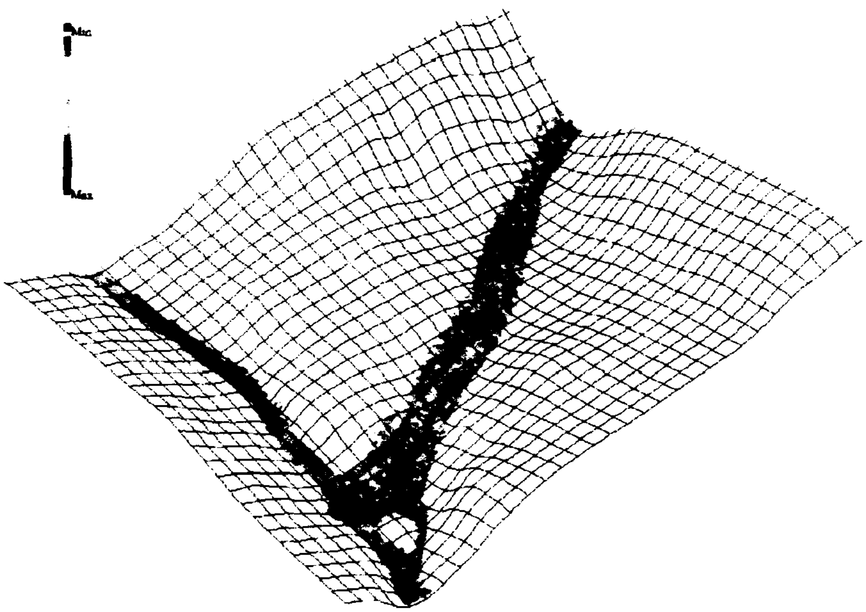


Figure 6. Grainsize Composition of confluence section.

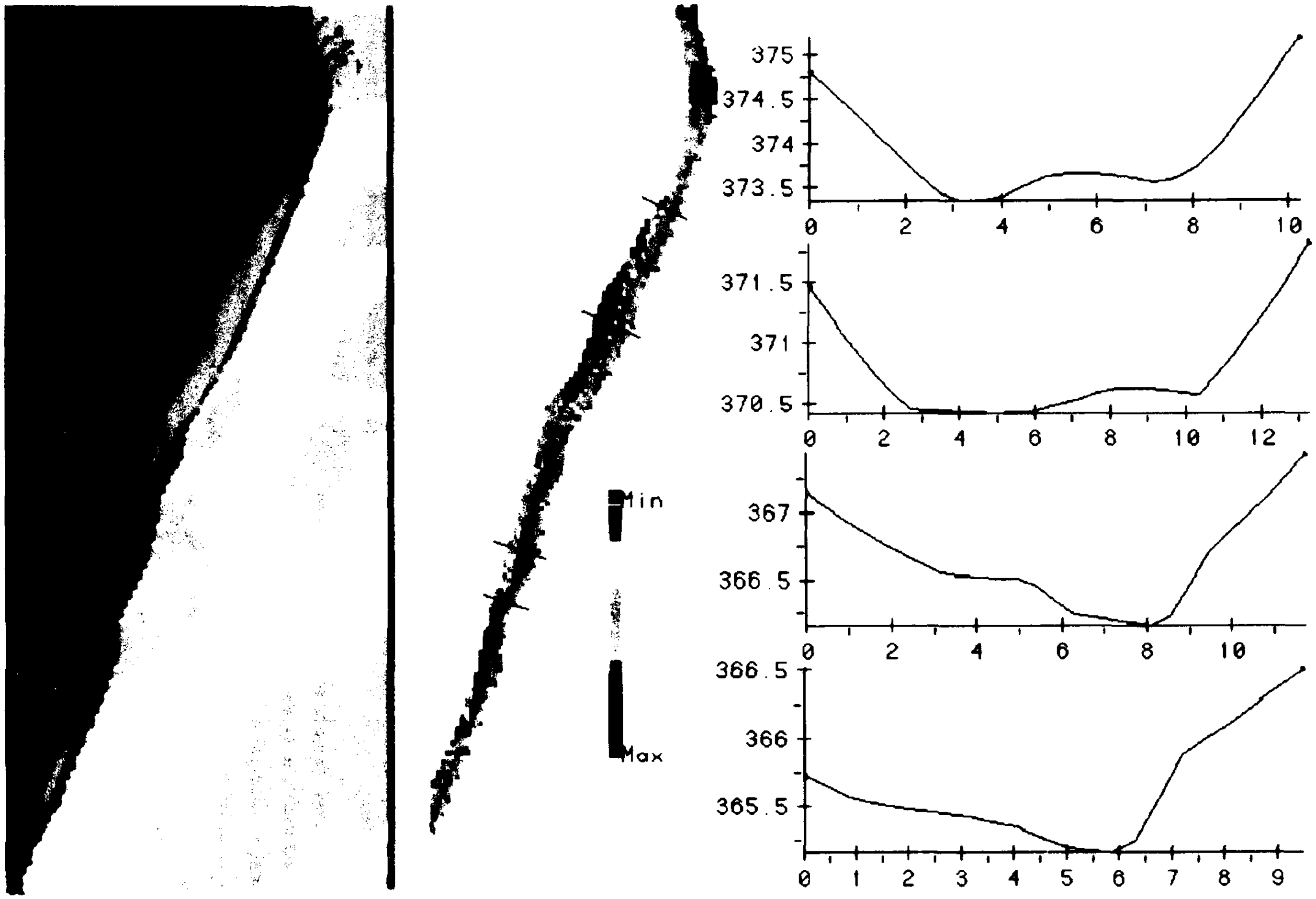


Figure 8. Section 2, showing plan view of shaded relief, Grainsize and four cross sections.

their planform and morphology. Although this is only a preliminary run of the model, a brief field recconaisance shows a high correlation in both location and morphology.

Discussion.

Observations of catchment dynamics show many examples of non linear behaviour, from the hydrograph output to sediment discharges (Lane *et al* 1997, Evans 1996). The model depicts similar behaviour, with an unpredictable sediment discharge, showing a partial decoupling between the hydrograph and sediment transport processes. Obviously there cannot be much sediment transport without a flood, but a flood does not pertain to sediment transport. The initial runs of the model, as described above, show the formation of landslides, berms, bars, braids, terraces and alluvial fans, of similar magnitude and form to those observed in the study area. These have all 'evolved' over the 15 floods, the model starting with featureless valley floors, equal initial conditions and distributions of sediment. The behaviour and formation of these features is all symptomatic of non linear behaviour. The grainsize distribution in figure 6 is a good example, with fines in areas of lower slopes where sediment has collected and armouring in the channels. Throughout the 16 runs of the model, there is a constant interaction between the channel and these stores, being re-mobilised and dispersed on some floods, yet left on others. The braided patter observed in figures 4 and 5 again is a result of these nonlinearities. The planform is constantly shifting, channels growing in one area, yet declining in another.

The model shows chaotic tendencies in its sensitivity to initial conditions. When the elevation data is saved to file, the values are truncated to 6 decimal places. When the data is re-loaded and the model run, different results emerge from when the values are retained in the computer memory at their full length.

Are these complex responses simply a condition of the models design? What happens to this response if more processes are integrated, such as a better hydrological model, or slope representation? Initial sensitivity testing hints that whilst altering the laws used gives different results, they are very similar. For example with figure 7, if this is run with a different sediment transport law, the exact dimensions of the berm / terrace sections is different, but their form and location is the same. Computational instabilities could explain non linear outputs, but to maintain stability, the amount eroded or deposited between each cell is limited to within a few percent of the local slope.

The implications of a model generating such a non linear response are considerable. We cannot rely upon a simple regression style model, because the response of the system is complex. The spatially distributed nature of the system means that we have to account for processes throughout the catchment. It is not the 'random' input from weather systems that is solely responsible for the non-linear behaviour of our fluvial systems, there is an inherent chaotic instability within the whole system. This is further demonstrated by the models sensitivity to initial conditions. Unfortunately, most fluvial modelling schemes, fail to account for non linear behaviour in any form.

If a catchments behaviour is unstable, sensitive to small perturbations in initial conditions, how can we incorporate changes that are so small to appear inconsequential, yet may prove to be important? Paola (1996) treats a 'whole' braided river system as a stochastic one, and finds the addition of a random element contributes to the accuracy of estimates of total flow and sediment flux. However, a chaotic system whilst appearing to give stochastic response is in fact deterministic. The LAB (Bridge & Leeder 1979)model of alluvial architecture is driven by an avulsion frequency, derived from a probability distribution around an observed mean. Whilst there are many other limitations to their approach (Heller & Paola 1996) similar approximations may represent one answer. Another approach may take the form of an AI answer, such as a fuzzy logic application or 'training' a neural net to incorporate this chaotic element. However, we may never get a true deterministic answer, having to rely upon an average of model runs, as climate modellers do.

The model highlights the importance of mass movement and slope processes in the evolution of a small catchment.

Disabling the landslide module resulted in a partial reduction in the non linearity. This suggests that the input from channel banks and heads is important, both as a sediment input and trigger for erosion/deposition episodes. Analysis of cut fill sequences, shows stream heads to be major contributing areas, as producers of sediment. This re-inforces research claims by Kirkby (1994) regarding the importance of the stream head in a networks evolution. There is still some non linear sediment response even when the mass movement section is removed, and this demonstrates the re-mobilisation and dispersal of sediment through out the catchment is also an important aspect of the systems behaviour. For example, the deposition of a clast may result in the lateral migration of the channel towards a pre-existing deposit, re-mobilising fresh material. In contrast to these positive feedbacks, there are several negative ones, controlling or pacifying the models operation. For example at the base of figure 7, where the channel has cut a terrace, incision is resulting in a stable channel pattern.

By choosing the 1m ^{2} scale, the effects of catchment scale processes such as hydrology and slope processes can be studied, as well as incorporating smaller scale catchment dynamics such as the in channel storage and re-mobilisation of sediment. This provides a clear advantage over models in which separate slope and channel modules are coupled together. With these schemes, different spatial and time scales have to be resolved and feedback's have to be explicitly defined. Furthermore, by selecting a 'meso' scale, this model demonstrates synergistic behaviour, showing that the overall catchment behaviour cannot be simulated simply from the sum of its individual component processes.

Conclusions.

Non linearities in catchment systems are crucially important at all scales, and we will never be able to fully account for all of them. It is not practical for large basin scale models to simulate three dimensional flow around clasts, yet the broader impact of such small scales must be incorporated. Similarly, three dimensional coupled flow and sediment transport models will have to account for irregularities in the time and space distribution of the arrival of sediment from upstream. Ultimately, the accurate incorporation of such factors will determine the power of our next generation of geomorphological models. Given the increases in computer power and advances in modelling techniques, it may prove that these 'chaotic' terms are the most important.

References.

Bates, P. D, Anderson, M. G, Hervouet, J. M-, and Hawkes, J. C. 1997. ' Investigating the behaviour of two-dimensional finite element models of compound channel flow'. ESPL, Vol. 22, pp 3-17.

Beven, K. J. & Kirkby, M. J. 1979. 'A physically-based variable contributing area model of basin hydrology'. Hydrological Sciences Bulletin, Vol 24, pp 43-69.

Bridge, J. S. & Leeder, M. R. 1979. ' A simulation model of alluvial stratigraphy'. Sedimentology, v.26, p. 617-644.

Coulthard, T. J. Kirkby, M. J. & Macklin. M. G. 1996. 'A cellular automaton landscape evolution model'. Proceedings of 1st International Geocomputation conference. Vol. 1, pp 168-185.

Coulthard, T. J. Kirkby, M. J. & Macklin. M. G. 1997. 'Modelling the impacts of Holocene environmental change on the fluvial and hillslope morphology of an upland landscape, using a cellular automaton approach'. Fluvial processes and Environmental change. (In Press) J. Wiley & Sons.

Desmet, P. J. J & Govers. G. 1996. 'Comparison of routing algorithms for digital elevation models and their implications for predicting ephemeral gullies. ' Int. J. Geographical Information systems. 1996, vol. 10, no 3, pp. 311-331.

Einstein, H. A. 1950. ' The bed-load function for sediment transport on open channel flows'. Tech. Bull. No. 1026, USDA, Soil Conservation Service, 71pp.

Evans, R. 1996. ' Hydrological Impact of a High-Magnitude Rainfall Event'. Advances in Hillslope Processes, Volume

I. Edited by M. G. Anderson and S. M. Brooks. John Wiley & Sons Ltd. pp 98-127.

Heller, P. L. & Paola, C. 1996.' Downstream changes in alluvial architecture: An exploration of controls on channel-stacking patterns'. Journal of Sedimentary Research, Vol 66, No. 2. p. 297-306.

Hoey, T & Ferguson, R. 1994.' Numerical simulation of downstream fining by selective transport in gravel bed rivers: Model development and illustration'. Water resources research, Vol. 30, No. 7, pp 2251-2260.

Howard, A. 1992.' Modelling channel migration and floodplain sedimentation'. in Lowland floodplain rivers: Geomorphological perspectives, J. Wiley & sons, 1-40.

Howard, A. 1994.' A detachment limited model of drainage basin evolution'. Water resources research, Vol. 30, No. 7, pp 2261-2285.

Kirkby, M. J. 1994.' Thresholds and instability in stream head hollows: A model of magnitude and frequency for wash processes'. Process models and theoretical geomorphology. Edited by M. J. Kirkby. John Wiley & Sons Ltd. pp 295 - 315.

Lane, S. N. & Richards, K. S. 1997.' Linking river channel form and process: Time, Space and Causality revisited'. ESPL, vol. 22. pp 249-260.

Macklin, M. G. and Lewin, J. 1993.' Holocene river alluviation in Britain'. Z. Geomorph. N. F. pp 109-122.

Murray, A. B. and Paola, C. 1994.' A cellular model of braided rivers'. Nature, 371, 54-57.

Nicholas, A. P., Ashworth, P. J., Kirkby, M. J., Macklin, M. G. and Murray, T. 1995.'Sediment slugs: large scale fluctuations in fluvial sediment transport rates and storage volumes'. Progress in physical geography 19,4, 500-519.

Nicholas, A. P. and Walling, D.E. 1997.' Modelling flood hydraulics and overbank deposition on river floodplains'. ESPL, Vol. 22, pp 59-77.

Paola, C. 1996.' Incoherent structure: Turbulence as a metaphor for stream braiding'. Coherent Flow structures in open channels. J. Wiley & Sons ltd. pp 705-725.

Parker, G. 1990.' Surface based bedload transport relation for gravel rivers'. Journal of Hydraulic Research, Vol. 28, No. 4, pp 417-436.

Polarski, M. 1997.' Distributed Rainfall-Runoff Model Incorporating Channel Extension and Gridded Digital Maps'. Hydrological Processes, Vol. 11. pp 1-11.

Quinn, P. F. and Beven, K. J. 1993.' Spatial and temporal predictions of soil moisture dynamics, runoff, variable source areas and evapotranspitation for Plynlimon, Mid Wales'.

Distributing Geographical Information Systems and Data Using Java and the Internet

David R Hill
Institute of Hydrology, Crowmarsh Gifford,
Wallingford, Oxfordshire, OX10 8BB. UK
+44 (0) 1491 692461
E-mail. drh@ua.nwl.ac.uk

Presented at the second annual conference of GeoComputation '97 & SIRC '97, University of Otago, New Zealand, 26-29 August 1997

Keywords

query	environmental	time series
spatial	4-dimensional	data model
database	SQL	GIS
Java	CGI	WWW
RDBMS	Internet	distributed

1.0 Introduction

Developing efficient and effective storage and access methods for large environmental databases is one of the main research aims of the Data and Software Systems group based at the Institute of Hydrology (IH).

The Institute of Hydrology investigates the effects of land-use, climate, topography and geology on the volume and character of water resources. It focuses on understanding water and energy fluxes arising from processes such as evaporation, interception and infiltration and modelling the hydrological cycle and chemical processes above and below ground. It is the aim of the Data and Software Systems group in conjunction with the scientists to design and implement software products for the dissemination of IH science. Many of these products involve the design and use of databases which are also used to manage IH's own environmental datasets. Much of the fundamental research behind IH's database designs took place in the period 1974 to 1990 during which time many of the commercial GIS packages which are in use today were not available or unable to deal with many of the problems presented by environmental datasets. As commercial GIS packages developed throughout the 1990's, research and development at IH moved to concentrate on environmental database design. There have been two key problems that IH has sought to ameliorate. One is that at present different data types are held in different systems making if difficult to explore relationships that span the different data types. The other is that the demand for data exceeds the IH Data Centre capacity to supply them. This paper will elaborate the problems and describe the underlying concepts involved in their solutions. It will then propose some suggestions for providing a simple query interface for environmental databases, that can be made available to remote users anywhere. The points made will be illustrated by reference to the IH's work on the Land Ocean Interaction Study (LOIS) (NERC, 1992).

2.0 Environmental Database Management

To improve understanding of coastal zone processes, NERC has invested over £25M in the LOIS programme (NERC, 1994). LOIS is a multi-disciplinary programme to study the movements of chemicals and then fluxes from the land into the rivers, out through estuaries and finally to the continental shelf and beyond. Information is vital to such a programme; the effective collation and manipulation of data from a wide variety of sources and subject areas being one of the keys to attaining the programme's scientific objectives.

2.1 Data Requirements for large thematic programmes

In order to manage effectively such a large data operation, the LOIS programme managers set up a data infrastructure. This infrastructure consisted of a Data Steering Committee and five Data Centres responsible for the acquisition and distribution of data from and to the researchers. The Rivers Data Centre, based at IH, is one of the five Data Centres and is responsible for acquiring, storing and distributing river based datasets. It is here that the systems described in this paper are being developed.

The diversity of the datasets that the database must accommodate creates a major challenge in terms of database design. These datasets include data that vary both in space and time ranging from river flow, water chemistry, species distributions, digital elevation data and river networks through to satellite images. It is the collective aim of the Data Centres to design and implement a unified database or environmental information system which is capable of bringing together these diverse datasets within one holistic database system. The hope is that by grouping all of these datasets within one integrated system, the task of researchers developing complex environmental models which cross component boundaries will be eased. This philosophy is supported by T.J.Browne (1995) who suggests that for an information system to be successful it must be holistic and interdisciplinary in approach.

2.3 Data acquisition and dissemination

Supplying and managing data for such a large thematic programme presents numerous problems for Data Centre managers, whose objectives are:

- To acquire major datasets from within and without NERC and make them available to the LOIS community.
- To establish standards for data definition and exchange formats.
- To provide data management services for LOIS data.
- To ensure long term security of the LOIS data and their availability to future science projects.

Traditionally, researchers obtain data from a Data Centre by writing, telephoning, E-mailing or completing a form on the Internet detailing the data that they require. The Data Centre then processes the data request and retrieves the data from the database. This process often incurs a delay as data requests may not be serviced immediately, however the formulation and execution of the query in the current system must be performed by Data Centre staff. Once the data have been retrieved they can be supplied to the user either by E-mail, FTP or the postal system. What both the Data Centres and scientists would like is the ability to browse and retrieve data remotely via the Internet. The acquisition of data suffers from similar problems. Presently, data arrive on many different forms of media and in many different formats. At each stage in the movement and translation process there are opportunities for data loss, corruption and delay. Advances in databases, networking and computer technology are now enabling these processes to be undertaken on the Internet and this will be discussed in the second part of the paper.

Before such an Internet solution can be designed, a clear view is required as to how the user can browse such a diverse array of datasets. It is a fair assumption that many users browsing the database will not have a detailed understanding of either the system or the types of data held within it. It is also likely that support will be minimal and that they will not want to master different methods of interrogation for each data type. Therefore, it would be an advantage if the user can perceive all data, whatever their type, to be held in one simple logical structure. Such a solution has been explored in the Water Information System (WIS) as described below.

3.0 Environmental Information Systems

To achieve data integration for the LOIS programme the Rivers Data Centre at IH is using the Water Information System (Tindall and Moore 1997; Moore 1997). WIS is an environmental information system which was designed and developed at the Institute with the backing of International Computers Ltd. Essentially WIS is a conceptually simple data model capable of storing generic types of data (Hill and Bellamy, 1996). It is implemented in a Relational Database Management System (RDBMS) on top of which sits a

UNIX based user interface. This provides an interactive geographical front end enabling users to visualise their data. The current implementation of WIS requires the data model to be implemented in ORACLE and the user interface software operates on a Sun workstation running SunOS 4.1.X. Both the hardware and operating system are now elderly and present various problems in terms of continued hardware and software maintenance. Much has been learnt since the initial development of WIS and what follows describes both the existing system and the improvements currently being implemented. However, the core of the system, the database design, has survived the test of time with only minimal modifications.

4.0 Database design

The WIS database design to which the system owes its immense flexibility, is best described in two parts, firstly the logical database design and secondly the physical database design.

4.1 Logical Database Design

4.1.1 Conceptual view of the data model

The logical database design provides a simple conceptual model which helps users to visualise how their data are stored. It allows the user to record the history of any object, or feature as it moves through space and time (Moore and Tindall, 1992). Descriptions of features and the events observed at them are recorded in terms of variables, parameters or determinands, known collectively in WIS as attributes. Thus, to store river water quality data, an individual monitoring site might be classified as a feature and the variables which describe or are observed at the site, such as its position, the site name, a unique reference number, river flow, pH values and so on, would be its attributes. Other examples of features could include roads, urban areas, maps, sewage works, licences and satellite images. WIS supports a wide range of spatial and non-spatial data types allowing the user to record most types of attributes. Examples of LOIS attributes could include names, reference codes, colours, centre lines, boundaries, soil types, the concentration of mercury and temperature. Both features and attributes are decided and defined by the users and their system and user definitions are stored in data dictionaries.

All attributes are assumed to be potentially time variant so even positional attributes may form a time series. For example, although a land based river monitoring station has a grid reference that is unlikely to change, marine and airborne sampling campaigns are conducted from a base that is constantly moving.

4.1.2 The WIS Cube

The description above provides one view of the logical design of the database. An alternative view of the same data is to imagine a cube of individual cells, as shown in figure 1.

The three axes of the cube represent features (where observations are made), attributes (what has been observed) and times (when the observations are made). Each cell contains a value (or values depending on the attribute's data type) of an attribute describing a feature at some moment in time. For example one cell might contain a real value representing the rate of flow in the river Thames at Teddington on the 20th May 1997. There are no constraints on the number of features, attributes or occasions which can be stored by the cube other than that imposed by the physical limits of the hardware. Listed below are the key properties of the WIS cube:

Any attribute may be observed at any feature;

A feature may have any number of attributes;

Any number of values may be recorded for an attribute over time at a feature;

The values may be recorded at fixed or random time intervals;

The data model does not distinguish between spatial and temporal data;

The Cube is infinite in all directions;

The significance of the cube is that it provides a completely generic data independent structure around which to build equally generic tools for data load, retrieval and analysis.

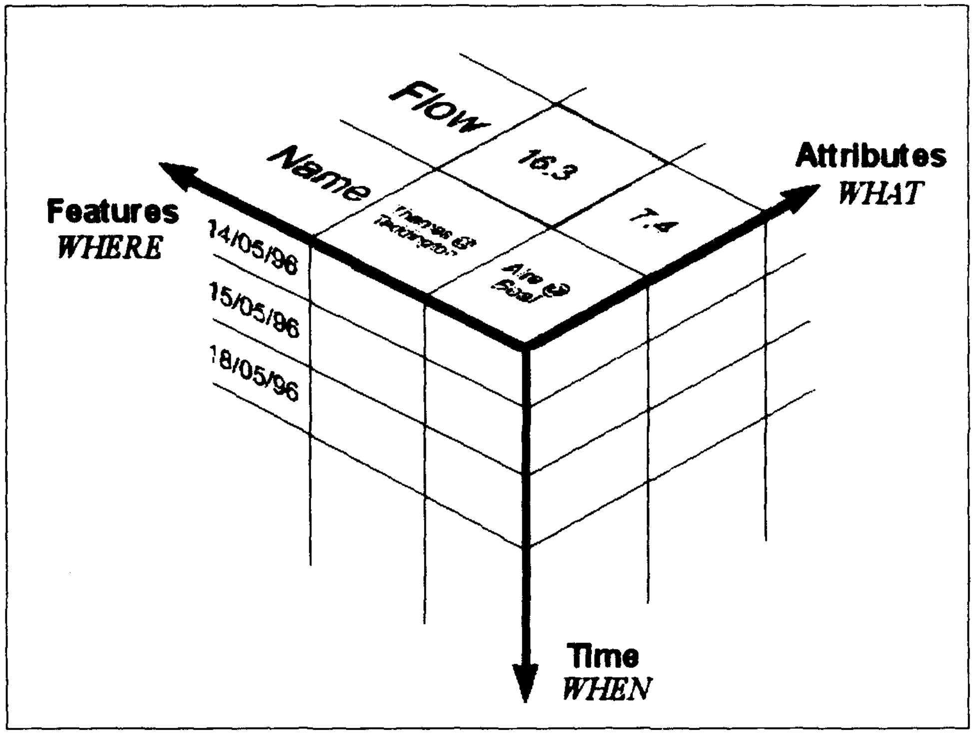


Figure 1 The WIS logical data model. The WIS cube.

4.2 Physical database design

The WIS cube can be implemented in any RDBMS. The underlying tables contain three different types of data;

Data Tables

The data tables share a common basic design and differ only in that different data types need different columns to hold the values. Each row in a data table contains a value from a cell in the cube. The first three columns of a data table contain the cube co-ordinates of the value, for example every value will have a feature ID (FID), an attribute ID (DID) and time ID (TID). For simple data types a fourth column holds the value. Thus, the DT_REAL table contains real values stored in a column called RVAL. Integer values are stored in the DT_INTEGER table in a column called IVAL. However, more complicated data types such as points are stored in the DT_POINT table and require three columns called X,Y and Z to represent their values. Other data types include names, character data, line data, grid data and binary (OLE) objects. Binary objects can be used to store JPEG images or, for example, Microsoft Word documents within a cell of the cube.

List Tables

The WIS search and select model relies on the concept of lists. A list contains a set of feature identifiers, attribute identifiers or date/time ranges which pick out the data required from the cube.

For example a 'where' list contains a set of features of interest. A 'what' list contains a list of attributes of interest. A 'when' list would contain a subset of the time axis. Combinations of what, where and when lists are also possible as in a 'where/when' list. An individual list is created by constructing the equivalent of a 'where' clause in a Structured Query Language (SQL) query. The range of logical operators, however, is greater and includes spatial operators and a facility to exploit parent/child relationships between features. Complex queries are possible by using

set operators on lists, such as UNION, MINUS and INTERSECTION. Sophisticated facilities for time matching are included that allow for the fact that values relate to different periods; some referring to an instant, others a day and yet others to a month or year. For example, a problem in the past has been that rainfall data are attached to rain gauges and river flow data are attached to gauging stations. Selecting flow data for occasions when it was raining was difficult if not impossible on most systems. The list approach allows the construction of 'when' lists of occasions when it was raining, which may then be used to extract flow data.

Reference data

Reference data can be divided into;

Standard data. Examples of these include units of measurement, methods, periods, statistics, methods, qualifiers and validation status codes.

Field and structure definitions;

These are the definitions of the data types that the system supports.

Feature type definitions;

A feature type is the primary classification of a feature and is the only mandatory attribute in the WIS data model.

Attribute definitions;

Attribute definitions comprise both the system and user information. The user information comprises of an identifying code, name, definition and reference. The system data include its datatype (structure), period, statistic and internal identifier.

Most users are completely unaware of the physical implementation of the database. However, application writers, programmers and modellers often want to interface directly with the database at a low level. To make this possible an object orientated database Application Programming Interface (API) is being designed and is currently being implemented for the latest version of the data model. The database API will provide the main access route to the database at a programming level. It has two roles: to make access easy and to protect the database from corruption. Generic data models are nearly always more difficult to query than specific ones. The API allows the user to express requests in a convenient way and then generates the SQL to answer them. The aim behind the API is to allow the programmer to think of the 'cube' as though it is a 3D array in memory. Assigning and using values in the database will be achieved by simple arithmetic statements. For example, a programmer could retrieve a value or update a value using the following statements;

```
value = mydatabase.cell(FID,DID,TID)
or
mydatabase.cell(FID,DID,TID) = value
```

Self evidently, rigorous validation checks for data that might corrupt the database will be included.

5.0 Distributing GIS and Data via the Internet

5.1 A changing computing paradigm

The original purpose of the generic data model was to facilitate the exploitation of relationships that span data types and to avoid the need to redesign the system whenever a new data type was introduced. However, a generic data model is also an important component in enabling remote data access to data. The WIS data model and database API as detailed in previous sections is used to form the core of a distributed GIS. As suggested in section 2.3, both Data Centres and scientists would like the ability to browse and retrieve data remotely via the Internet. Up until recently computing technology has not been able to allow the development of such systems. However, changes in this situation will soon mean connecting to the Internet will be as common as using the telephone, resulting in a web browser on virtually every desktop computer. It enables simple communications between millions of people throughout the world from a common user interface. The software which will operate on these browsers could be written in the Java programming language (Sun Microsystems, 1997). Java was designed to provide a platform and operating system independent programming environment. Although Java is relatively immature in terms of computer languages, it provides some fundamental advantages over its rivals which can be summarised as follows;

- Applications or applets may be written once and executed on any platform, reducing development costs.
- Applications or applets may be downloaded on demand from a centrally administered server.
- Java provides the advantages associated with object orientated languages.
- Java has been designed for communication across the Internet therefore security issues have been properly addressed.
- Java removes the programmer from the complexity of pointers and memory management found in languages such as C/C++.

In many cases the actual Java language itself is not the most important development but rather the introduction of the Java Virtual Machine (JVM). Java computing operates in a client/server environment where applets are dynamically downloaded on demand from a server.

Figure 2 illustrates how this computing paradigm can provide a solution for distributing the means of querying a database and subsequently viewing and retrieving data. This methodology can of course be used for the reverse process of submitting data to Data Centres. Imagine the following situation; a user goes to his client terminal, it could be a PC, Unix Workstation or Network computer, and connects to the Data Centre Web page. From this page the user indicates that they would like to browse the database. The Web server deals with this request by sending the appropriate Java applet to the client machine which loads and runs the applet in memory. For a LOIS scientist, the applet might enable the user to formulate questions to the database. The questions are first sent to the Web server where the Java applet makes a database API call. The database API calls manage the connection, the formulation and execution of any user queries. Internally the database API calls produce SQL queries which execute on the database server. Any result produced by an API call is then sent back to the user in either textual, data or graphical form or as a direct data input stream for a Java applet already running on the client machine. As far as the user is concerned, any connections are established directly between the user and the database server as indicated by the loop drawn of figure 2. However, communication between the user and the database is managed transparently by the Web server which supplies Java Applets or Web pages on demand.

Java applets can be as simple or a sophisticated as desired. For the LOIS programme it is hoped that they will enable the users to query the database with the aid of simple maps and have any results presented as reports or graphs. The results of queries should also be available as a simple

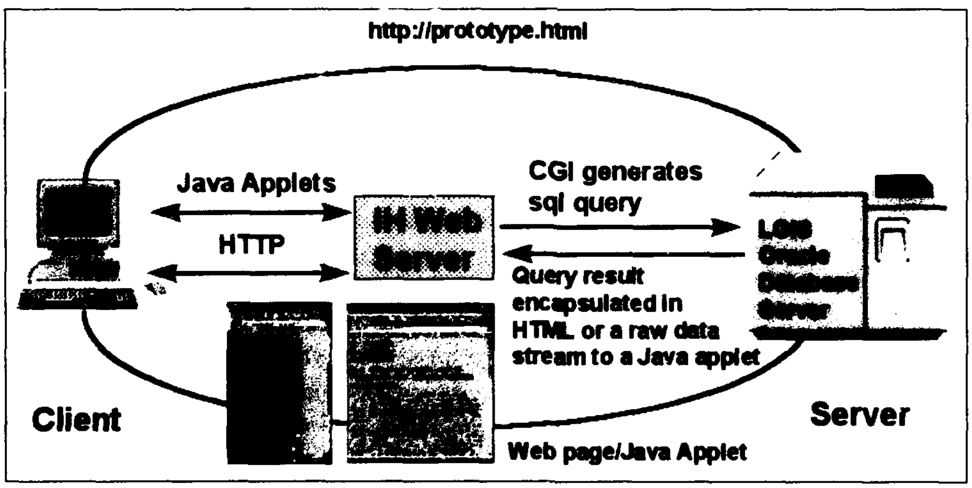


Figure 2. Distributing GIS and Data Using the Internet and Java.

export file that can then be imported, for example, into a spreadsheet for analysis. The Java applets would provide the basic GIS functions such as pan, zoom, scale and re-projections. More complex GIS functionality such a river climbing and catchment boundary derivation could be coded in Java or Common Gateway Interface scripts depending on the processing requirements. However, access by this means is currently only feasible if the number of different database systems that must be queried can be kept to a sensible minimum, ideally one. Hence, the desire for a single all purpose data model.

5.2 Advantages and disadvantages of distributing GIS and data

Distributing GIS capability and data has many advantages for both the user and the Data Centres. Obvious benefits for the user include a common single interface to a large comprehensive data source. Users would also have the ability to express spatial and time series queries from a map based user interface and be presented with the option of downloading the results for further analysis.

Java operates in a client/server environment enabling system developers to determine where processing is undertaken. For example, Data Centres do not want the computing overhead of executing and maintaining the display of the clients user interface. Java allows the applet to be downloaded onto the client machine and executed on their local processor. The only processing carried out by the Data Centre is the preparation and execution of the user's query.

Java is a relatively young language and many of the techniques described in this paper have yet to be tested. Much of the Java work, is however, in the initial stages of design and prototyping. Problems have occurred when attempting to establish large data stream connections with remote database servers. Many scientists have expressed concerns about the security of their intellectual property rights with regards to their datasets. Java does have a security model which has been designed for Internet use. However it is still unknown exactly how secure this model is and comprehensive tests will need to be undertaken before releasing any system to the public.

6.0 Conclusions

The WIS data model described in this paper has illustrated that it is possible to combine many diverse spatial and temporal datasets within one physical database and thus facilitate the exploration of relationships that span different data types. However, what is now required is an API that allows modellers to interact easily with the database. To achieve this an object orientated approach is being adopted that represents the data as composing of three object types, a database, dataset and cube cells. The data values are the properties of these objects and associated methods allow their manipulation.

The paper has attempted to provide an insight into the future developments of environmental information systems and the way in which the Internet will influence the design of such systems. Java and other associated Internet technologies have provided system developers with a rich set of tools and protocols for developing distributed systems. However, the success of Java may not be entirely due to the language itself but the introduction of the JVM. There have, however, been suggestions from a leading hardware vendor of developing a Universal Virtual Machine which would be capable of producing byte code from any of the main stream computing languages such a C/C++ or Visual Basic. Should this come about then the need for Java could evaporate.

Distributing simple GIS capabilities via the Internet has many advantages for users and Data Centres. Firstly, the task of browsing and retrieving data from the database becomes the responsibility of the user. Users may also download the results of their queries at their convenience, for further analysis. By moving the onus for browsing and retrieving data onto the user, Data Centres then become free to investigate other problems such as quality control, quality assurance, security and visualisation techniques.

7.0 References

Browne, T. J. (1995) The Role of Geographical Information Systems in Hydrology. *Sediment and Water Quality*

in *River Catchments*, UK, 1995, John Wiley & Sons Ltd, Pages 33 - 48.

Hill, D. R. and Bellamy, S. P. (1996) Search mechanisms for querying the time dimension in 4-D GIS. *1st International Conference on GeoComputation*, Leeds, UK, 1996, University of Leeds, Vol 1, Pages 405 - 420.

Moore, R.V. (1997). The logical and physical design of The Land Ocean Interaction Study database. *The Science of the Total Environment*, UK, 1997, Elsevier Science, Pages 137 - 146.

Moore, R.V. and Tindall, C. I. (1992) What is WIS? *IH/ICL Report*, Wallingford, UK

Natural Environmental Research Council (1992) Land-Ocean Interaction Study (LOIS) Science Plan for Community Research Project. *NERC*, Swindon, UK.

Natural Environmental Research Council (1994) Land-Ocean Interaction Study (LOIS) Implementation Plan for a Community Research Project. *NERC*, Swindon, UK.

Sun Microsystems (1997) The Source for Java, 1997, *Sun Microsystems Inc*, http://java.sun.com/

Tindall, C. I. and Moore, R.V. (1997) The Rivers Database and the overall data management for the Land Ocean Interaction Study programme. *The Science of the Total Environment*, UK, 1997, Elsevier Science, Pages 129 - 135.

Using Middleware to Provide Geographic Information Systems (GIS) Based on Relational Databases

Andrew Cardno, Kelly Buchanan & Dr Pete Frizzell
PO Box 4103, WELLINGTON
Phone [64] (4) 472-8244
eMail: andrew@critchlow.co.nz
or kelly@critchlow.co.nz
or peter@critchlow.co.nz

Presented at the second annual conference of GeoComputation '97 & SIRC '97, University of Otago, New Zealand, 26-29 August 1997

Abstract

We describe the process of using existing database middleware tools for deployment of geographic information systems across the enterprise. We will investigate the impacts on system design, including how the relational model can be simplified by incorporating implicit geographic relationships. We will also discuss how the multi-tiered environment impacts on deployment.

Data integration and sharing using SpatialWare will be investigated showing how it allows one to manage spatial and business information in a single database. This allows for widespread sharing of data while eliminating much of the expense associated with data duplication and local storage of data. We will discuss how data integration is transforming spatial information systems, allowing existing IT infrastructures to manage their geographic data.

We will describe how by conducting much of the spatial analysis on the server, SpatialWare delivers more efficient and productive processing of spatial queries. It will be shown how a spatial server provides distributed processing of spatial queries using a standard relational database as the repository for the data. In addition we will discuss how this results in reduced network traffic, improved response time and thin spatially enabled clients, when compared to traditional GIS.

1. Introduction

Middleware spatial tools provide two broad areas of integration:

They allow spatial data to be directly stored in the relational database.

They provide a layer for query on this spatial information.

Integration of the middleware components is as either: part of the server architecture, as with Informix datablades; or a separate process where the server has no knowledge about the spatial server. Middleware layers can be held physically on the server or run on a separate server.

Spatialware is a middleware layer which can store its data directly in Oracle or Informix. This builds on the core functionality of the server, thus allowing normal data management principals to be applied across both spatial and non-spatial data.

Distribution across the enterprise may entail multiple spatial servers, which distributes the processing load, giving true scalability and deployment across the whole enterprise. Traditional database clients can continue to have access to the database server in parallel with the spatial users, while end users applications can be spatially enabled to provide spatial query in a thin client situation. Multi-tiered distribution of spatial information servers allows

load distribution across the enterprise to provide rapid response to spatial queries.

The first part of this paper will explore important issues regarding the use of middleware to provide GIS solutions based on relational databases. We will consider impacts on system design, including how the relational model can be simplified by incorporating implicit geographic relationships. The issue of data integration will then be addressed, followed by a description of how middleware can assist in providing a thin client.

A description will be given of how SQL has been extended with spatial functions and predicates that let users perform all the typical spatial data modelling and analysis capabilities required by GIS systems.

We will then examine how the extended SQL that SpatialWare uses is open, with published standards for interfaces, data storage and operations, in particular the *May 1996, SQL/Multimedia and Application Packages (SQL/MM) standards* for spatial data handling. This means that users can access and manipulate mapping information using the common programming language used throughout the database world. Following this, we will discuss how this standard provides for abstract data types to collect point, line and polygon geometric primitives into instances of a spatial object. In addition, we will outline how SQL/MM has been implemented, defining the abstract data type.

The second part of this paper investigates the importance of data in spatial data warehouses, with considerations of integrity and management. We will discuss the implications of business rules including topological constraints, and will show how organisations can use spatial business rules, without mapping components, to improve analytical and operational parameters.

We will conclude that data and data management is the most important component of spatial systems. Traditional IT infrastructures will use new middleware tools for geographic data management in relational databases.

2. Important Issues on the Use of Middleware to Provide GIS Based on Relational Databases

2.1. Impacts on System Design

Implicit relationships between spatial entities result in vastly fewer explicit relational joins being necessary. For example, in modelling road centrelines, parcel boundaries are implicitly connected to the road. Thus, if a new road is constructed, then the relationship between the road and the parcel does not have to be explicitly stored in the database. While this simplifies the system design, it also creates implicit business rules. As these rules are not always obvious, careful consideration is needed in defining the data to ensure correct application. One example of a non intuitive interaction between seemingly unrelated features is the interaction between address and parks. If address were dynamically segmented along roads then parks would be considered as breaks in address, thus the address segmentation and park feature are related. As the relationship between park and address is implicit, the segmentation algorithm will need to take into account the park.

The following modelling techniques have proven to be useful in defining geographic entities that follow the correct behaviour and simplify business rules:

Abstraction to layers of interaction: In a traditional GIS, features tend to be represented in functional layers, such as a roads layer. As the database provides views of the data, one further layer of abstraction from a functional feature to an interactive feature allows business rules to be defined for super classes. Consider the dynamic segmentation model for representation of linear data along a road. One level of abstraction allows all segmented data to be stored in the same entity with subclasses defining attributes and the abstract class defining the segmentation. Hence, the interaction between segmentation data and the actual linear features is abstracted and business rules are simplified. This abstraction can also be called grouping of interactive features.

Separation of non interactive features: Inclusion of non interactive features with the same functional class should be avoided. For example, definition of parks and parcels should be in non-interactive layers. Often parks may be exactly defined by parcel boundaries, however there are many cases where the physical park boundary differs from the legal definition of the boundary. By modelling the park in a separate non interactive layer, dynamic queries may be used to extract the parcel definition of a park without losing the physical boundary definition.

Apply spatial checks to spatial entities: Often, business rules need to involve interactive checking processes. Again, consider the relationship between parks and addresses. Sometimes a park may have been allocated an address, or part of a park may have been allocated an address. These rules relate to the physical position of the park, as a building on the park may have an address. Definable business rules between parks and addresses needs to involve human intervention and checking to allow for these unusual exceptions.

2.2. Data Integration

Data integration is transforming spatial information systems, allowing existing IT infrastructures to manage their geographic data. Traditional spatial information required specialised knowledge of the proprietary system to execute day to day data management activities. By spatially enabling traditional databases, the existing business information systems can be built on to give additional spatial storage. This extends the traditional data warehouse allowing database administrators to view spatial data as just another database and as such requires no special skills to manage the data.

Within the corporate sector, there is considerable resistance by information technology managers to additional data repositories outside of their existing infrastructure. Data integration main-streams the concept of spatial data management, and implies buy in from IT managers who are resistant to the high risks involved in relying on specialised skills to manage part of the data repository.

The extension of the single database to give widespread sharing of corporate data can lever off the existing IT infrastructure allowing enterprise wide access to the spatial data while avoiding duplication of data and local storage issues.

2.3. Thin Clients

Middleware layers provide powerful spatial query results, passing only the results to the client. Spatial query servers like SpatialWare use spatial indexes to optimise the queries and minimise the impact on the server, while never downloading the spatial data to a separate spatial repository. These spatial indexes provide implicit relationships between spatial objects which can be used as joins and filters to produce sophisticated implicit analysis of the data. There are two broad groups of clients for spatial information servers, non graphical and graphical.

2.3.1. Non-graphical

Non graphical user interfaces can apply spatial business rules to analysis of data. These interfaces typically view small sections of the data, often one record, where this small piece of data may be the result of a complex spatial query. Spatial query servers execute the query and pass only the result back to the client. Minimal network bandwidth is required for this type of client.

2.3.2. Graphical

Graphical users view much larger sections of the data at a time. While the spatial queries are still processed server side, client side handles considerable large data throughput. Cartographic design can often be used to reduce the amount of network bandwidth needed to draw the map. These type of users represent a significant risk in terms of network bandwidth and server load.

When designing the clients it is important to manage the data requirements for map redraw. Vector based distribution of maps have a variable and possibly high data volume requirement. In the worst case, it is possible for a user to request all spatial information in the database to redraw one map.

Raster based distribution of maps across the corporation provides a manageable medium where the network load is definable and related to the number of users. These raster images can be given much of the function of a vector map by using spatial queries to drill down as the user interacts with the system. These raster maps can be generated dynamically in a separate middle-ware layer to provide enterprise-wide thin client with both spatial queries and mapping functions. This differs considerably from traditional GIS solutions which typically use vector map redraw and often require large local storage or high network bandwidth.

2.4. Spatial Extensions to Structured Query Language (SQL)

Spatially extended SQL provides an accepted interface to spatial data query. Organisations can extend their existing applications to provide spatial query by using standard SQL result sets via commercial tools such as Visual Basic, Delphi, PowerBuilder, or C++. This middleware enabling process gains industry credibility with the establishment of SQL standards relating to spatial query.

The major extensions to the SQL standard, as defined by Spatialware, include:

A spatial data type for storing spatial data

Additional SQL operators and predicates for manipulating spatial data

An object-oriented operator for sub-classing and inheritance

An extended transaction model that includes transactions involving spatial data

Additional data dictionary functionality to maintain the information required for projects, thematic display, transformations, and accuracy

SQL is extended to include the following spatial predicates:

Overlaps: Is used to determine whether the selected object overlaps another selected object, where and object can be a point, line or polygon.

Contains: Is used to determine whether the selected object completely contains another selected object.

Contained by: Is used to determine whether the selected object is completely contained by another selected object.

Adjacent: Is used to determine whether the selected object shares any points with another selected object.

At start of: Is used to determine whether the start point of a line is touching a selected object at just that point.

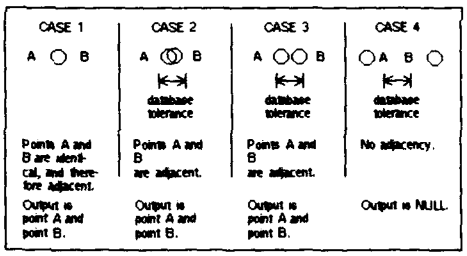
At end of: Is used to determine whether the end point of a line is touching a selected object at just that point.

Connected to: Is used to determine whether the start or end points of a line are touching a selected object at one of the start or end points.

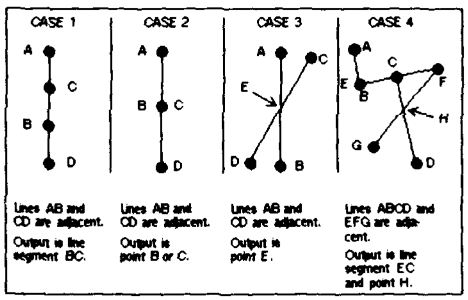
The following functions are included in spatial SQL:

Adjacent: Returns the common points of the two input objects. The following examples, drawn from SpatialWare help, detail how adjacency is utilised:

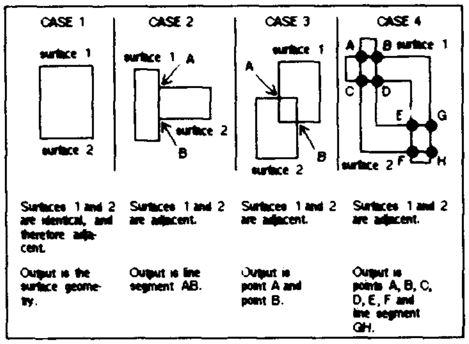
Two points have an adjacency if they have the same coordinates (which means they are identical), or are within the database tolerance of each other.



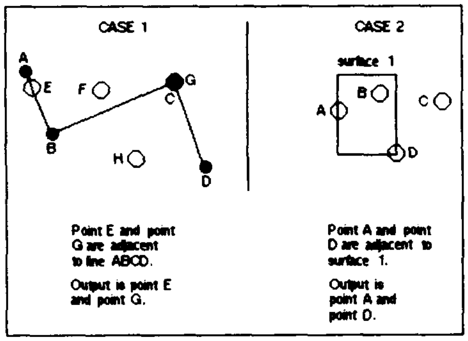
Two polylines have an adjacency if they share any line segment, or intersect at any point(s).



Two polygons are adjacent if their boundaries share any line segment, or their boundaries intersect at any point(s).



A point and a polyline are adjacent if they intersect at any point, and a point and a polygon are adjacent if the point intersects the polygon's boundary (or vice versa) at any point.



Minimum Enclosing Rectangle (MER): Returns the smallest rectangle that can contain the submitted object.

Buffer: Returns an object which encloses the specified object by the defined buffer distance.

Centroid: Returns the point which is the centroid of the specified object.

Overlap: Returns an object corresponding to the intersection of the two specified objects.

Contain: From two submitted objects, the first object is returned if it is completely contained by the second object.

Geometry_union: Returns the geometric union of the two specified objects.

Length: Returns the length of the selected object.

Slope: Returns the slope of the selected object.

Area: Returns the area of the selected object.

Perimeter: Returns the perimeter of the selected object.

Skeleton: Returns the skeleton of the specified object.

The following example displays the use of spatial SQL in a query:

```
select parcel.owner,
    area(overlap(buffer(road.sw_geometry, 100, 1),
     parcel.sw_geometry))
 from road, parcel
 where buffer(road.geometry, 33, 1) overlaps
    parcel.geometry
  and road.status = 'formed';
```

This selects the area of road buffered to 100m overlapped with the specified parcels, where a 33m buffer of the road overlaps the specified parcels.

The extended SQL defined in this section conforms to the *May 1996, SQL/Multimedia and Application Packages (SQL/MM) standards* for spatial data handling. The next section will discuss a specific Spatialware implementation of SQL MM for utilisation on an Oracle database.

2.5. Spatialware Implementation of SQL MM for Oracle

SpatialWare has implemented the SQL MM standard to provide a middleware layer for spatially enabling Oracle databases. Currently, this interface has been implemented on Oracle and is soon to be extended to Informix datablades. We will discuss the Oracle implementation.

Within the Oracle database, Spatialware adds two columns to each table to be spatially enabled. These two columns are named SW_MEMBER and SW_GEOMETRY. The SW_MEMBER column is of type 4 byte integer and is used by the spatial indexing. SW_GEOMETRY is of type long raw and holds a blob representation of the spatial object. As Oracle is limited to one long raw per table only one spatial object can be added per tuple. This limitation can

be overcome by using separate tables to store alternative spatial views. These two columns together represent the ST_SpatialObject container that holds the following geometric primitives: points, polylines, circular arcs and polygons with or without holes. In addition, and invisible to the non DBA user, are tables that hold the spatial data dictionary.

The spatial index tables provide the backbone needed to efficiently execute spatial queries while minimising the load on the Oracle server. The elegance of this solution is shown when considering the implications for consistency and security of the data. Spatialware uses the existing Oracle security defined for the table, hence, users' access rights to the spatial data and business data will match.

The spatial query performance under load matches the Oracle performance, so the SpatialWare middleware layer builds on the scalability of Oracle. Furthermore, separation of the middleware layer onto a distributed server accesses the Oracle server without competing for computer resource.

2.5.1. Query Optimization in SpatialWare

SpatialWare for Oracle does its own query parsing, which manages RTree and all other issues relating to the database. The SW_Optimizer calculates the most efficient access path, in essence deciding on the how queries are passed through to Oracle. If a query is non spatial, then it is passed directly to oracle, while spatial queries are resolved in a number of steps: if the query involves a join then a greedy join algorithm is used where the right hand side is visited once for each object on the left hand side.

2.5.2. Example of Query Optimization for SpatialWare

The following example has been taken from the SpatialWare help documentation.

```
SELECT C1.*, C2.* FROM C1, C2 WHERE C1.GEOMETRY
    OVERLAPS C2.GEOMETRY;
```

This query also joins all the attributes of class C1 with the attributes of class C2. However, in this case, there is a "spatial join predicate". The classes are still joined in the order specified by the FROM clause, but performance is improved because, for every feature/record in class C1, only the record with those features in C2 that satisfy the join predicate is joined - in this case, only those features that overlap the feature of C1.

By eliminating records from the inner plan through the use of a join predicate, performance is improved because fewer records need to be handled in total, i.e., rather than comparing every record in the first data set with every record in the second, only those records from the second data set which satisfy the join predicate are joined.

3. *Significance of Data in Spatial Data Warehouses*

Spatially enabled databases extend the scope of the existing corporate database enormously. Current business information can be spatially enabled to add real value. In addition, other data sets are implicitly spatially related, which extends the scope of the database enormously, providing analysis between data sets which are seemingly unrelated (until the spatial characteristics are considered). This consideration differs markedly from the traditional data matching requirements of databases where all data needs to be modelled and integrated.

Consider the addition of statistics New Zealand meshblock data. In a relational database, each of the entities that need to have a relationship represented will need to have a data matching process, followed by alteration of the entity to contain the meshblock relationship. Within a spatial data model, this relationship is implicit in the data, hence, all spatial entities are implicitly related. Correspondingly, a spatially enabled database needs to be focused on adding spatial datasets of relevance to the business.

3.1. Topological Constraints and Business Rules

Business rules in a relational database are strict and defined, and are generally managed dynamically as the data is altered. Within a spatially enabled database, these business rules are difficult to manage dynamically. Consider the relationship between meshblock and location. If a new set

of meshblock boundary tables are input, the boundaries may have moved significantly, and clients' meshblock may have altered. Effectively, this is altering the relationship with none of the relational referential considerations. An example business rule could be a checking process to verify whether all clients are in the same meshblock as previously defined, and producing an exception report that needs to be manually verified.

Critchlow Associates Limited have recently undergone this exact process with the introduction of the 1996 census meshblocks, which differ in location to the 1991 census as the boundaries have been refined to more closely match road centrelines.

3.2. Querying Without Mapping Components, to Improve Analytical and Operational Parameters

Spatially enabled entities are implicitly related to all other entities in the same world. Consider a simple database of spatial entities with no explicit relationships. All entities are related implicitly and much significant business information can be determined. A more specific example is the relationship between reported crimes and their police stations. These crime records are explicitly related to the station by the dispatch location. However, they are implicitly related to their station by location. A resourcing study of stations would be biased towards the stations with higher police resource showing how many incidents were acted on, while the spatial analysis would give a different picture of resource demand around each station. The most interesting analysis is a combination of both the spatial and the relational models. In the station/incident case, true allocation of resource could be considered in both a historical and a predictive sense, where the spatial information system produces predictive, and the relational historical.

This increase in the information content of the existing database, by spatially enabling produces better operational information. Implicit relationships do not pre-suppose modelling and thus can have unbiased consideration, while relational models enforce modelling, therefore implying bias. Consider station/crime example. If the dynamic operational resourcing requirements were allocated in real time then the relational model would only state who was doing the work. If inefficiently allocated resource enforced poor travel time decisions, then these decisions would be hidden resulting in increased operational resourcing. Alternatively a spatial operational system would show demand location; contain more information relating to travel time; provide the implicit relationship between station and resource requirement; and would produce improved operational parameters by avoiding pre-supposed business modelling.

4. Conclusion

Data and data management is the most important component of spatial systems. Much of the traditional resistance to spatial information systems has resulted from the proprietary data storage mediums and the need for specialist skills to manage spatial data. Spatial middleware layers remove this necessity, allowing traditional data warehouses to store and manage spatial information using relational database technology.

In addition to the ability to store the data, middleware layers enable spatial queries across the enterprise. Efficient businesses will use these queries to build spatial business models that allow greater efficiency. This efficiency comes at a significant innovation cost as many of the implications of spatial business are not obvious.

5. References

Cardno A.J. and Frizzell, P.W. (1996). "Integrating Successful Spatial Applications" *Proceeding of SIRC96 Conference, SIRC, University of Otago.*

Pacheco X. and Teixeira S. "Delphi Developer's Guide" *Borland Press.* Sams Publishing.

Oracle On-line Help Documentation.

SpatialWare On-line Help Documentation.

Todd B. and Kellen V. "Delphi: A Developers Guide" *M&T Books.* 115 West 18th Street, New York.

The Visualisation of Relationships Between Geographic Datasets

Mark Gahegan
Department of Geographic Information Science,
Curtin University, PO Box U 1987, Perth 6845,
Western Australia.
Tel: +618 9266 3309, fax: +618 9266 2819
email: mark@cs.curtin.edu.au
URL: http://www.cs.curtin.edu.au/~mark

Presented at the second annual conference of GeoComputation '97 & SIRC '97, University of Otago, New Zealand, 26-29 August 1997

Abstract

A collection of geographic data from a particular region contains many explicit and implicit relationships. The data will typically have been gathered according to different models of geographic space (Goodchild, 1992). Furthermore, low level data is often synthesised into higher level objects to which more meaning is ascribed (semantic abstraction). This paper addresses the problems of exploring such interconnections between data using state-of-the-art visualisation techniques and is based on the premise that visual exploratory data analysis is a useful tool for providing insight into the complex and subtle relationships that occur in geography (Tang, 1992, Gahegan, 1996). Results are given in the form of images (in the paper), VRML scenes and video clips (which may be downloaded from the web). The tools and techniques described extend beyond what can be currently achieved in commercial GIS in terms of (i) the flexibility of scene description, (ii) the volume of data (particularly the number of layers viewable concurrently), (iii) the amount of interaction available to the user and (iv) the facilities with which to study relationships.

1. Introduction

Currently available GIS are poorly equipped to visualise the complexity and volume of data that is routinely used in spatial analysis and modelling; this is rather ironic, given that the display of data is a primary function. Instead of extending an existing GIS (Hartmann, 1992), or building specific tools (Haslett et *al.* 1991), we have chosen to exploit existing visualisation environments, specifically IRIS Explorer and VRML-2 (ISO IEC, 1997). The reasons for this are that GIS have quite restricted graphical capabilities, which are usually difficult or impossible to customise. By contrast, current visualisation systems can support our needs well, albeit with some problems importing data.

The aim of exploratory visualisation is not to analyse the data *per se*, but rather to *present* the data to the user in a way that promotes the discovery of inherent structure and relationships. In psychometric colloquialism this is known as *inducing visual 'pop out'* (Csinger, 1992). Thus, a collaborative mode of interaction is developed between the user and the machine, where the visualisation environment produces a stimulus which is then *interpreted* by the user, enabling full advantage to be taken of the abilities of humans to perceive complex structural relationships.

The purpose of this paper is to concentrate on some specific techniques for studying relationships between layers of data. A justification for these techniques is given first, drawn from the relevant psychometric literature. Some examples of their use are then presented and discussed.

1.1 The State of the Art in Exploratory Visual Analysis

The use of virtual reality and visualisation tools to study interaction within and between datasets is a relatively new

idea (Hearnshaw & Unwin, 1994). Most of the research conducted to date is as yet only qualitative in nature, due to three distinct problems:

1. The difficulty in defining the psychometric principles that a 'good' visualisation should follow.
2. The difficulty in constructing visualisation environments which use these psychometric principles to good effect.
3. The difficulty in measuring quantitatively whether a specific visualisation follows these principles and validating the techniques using human subjects.

Bertin (1981), Mackinlay (1986) and Rheingans & Landreth (1995) address the first of these problems, providing useful guidelines from the science of visual perception. Upson (1991) gives a useful introductory account of the visualisation system design process. Freidell *et al.* (1992) and Duclos & Grave (1993) show how such an approach may be automated using rules and grammars. O'Brien *et al.* (1995) and Gahegan & O'Brien (1997) describe such a rule base, designed around the needs of geographic datasets. Jung (1996) goes on to show how effectiveness criteria might be applied to an automatically produced visualisation, to evaluate its usefulness. However, most of the work in this field to date addresses the first two points; it is perhaps as yet too early to address the third.

The visualisation paradigms adopted here are those of *axis* and *mark composition* (Senay & Ignatius, 1991; 1994) where a number of spatially referenced datasets are represented using a (usually smaller) number of concurrent surfaces and layers of symbols or icons (Gahegan, 1996). The idea is to provide a spatially compact 'stimulus space' within which visualisations may be constructed. The 'natural world' paradigm suggested by Robertson (1990) is utilised where possible so that the scenes have the look of a conventional landscape. This has been shown to work well with the human visual system, which is highly optimised to interpret the various 'landscape metaphors' (Rheingans & Landreth, 1995).

1.2 The Use of Interactors

By themselves, these paradigms have some cognitive limitations, caused by the necessity to separate data into different layers, to avoid over-cluttering in any one layer. When many layers of data are required, as is common with axis composition, then the user's focus of attention must 'shift' between layers in order to assess their inter-relationships (to 'see' pattern or structure). Attention shifting is undesirable; it leads to a weakening of the overall stimulus (Bertin, 1981) at any given point in space, since it is divided amongst *n* layers. In a cognitive sense it is therefore true to say that "*the whole is greater than the sum of the parts*".

A mechanism is required to establish links between layers of data. To be useful for exploratory analysis this mechanism must facilitate perception of the structural and positional relationships between specific regions in the data. The solution adopted here is the use of 'interactors' - graphical icons that emphasise positional information. A number of different geographical *interactors* are used to communicate a variety of types of relationships between data layers. Types of interaction include the projection of objects, pixels, lines and points between data layers to describe processes such as interpolation, object extraction, classification, edge detection and so forth. A library of interactors has been implemented by creating modules within the IRIS Explorer environment. Figure 1 shows an example screen shot (from Explorer) showing the specification of an object (polygon) interactor describing a geological region.

This paper principally describes one class of interactor, used to examine relationships between datasets, particularly between higher objects and the data from which they were made. Two specific approaches have been investigated, both of which involve the use of animation to communicate interaction. Animation techniques provide a powerful and visually effective means of studying the relationships between higher objects and their defining data (Keller & Keller, 1993). Movement has been shown to have a high visual impact, and its detection in humans uses significantly different neural pathways (Livingstone & Hubel, 1988) to the perception of 'retinal' variables; namely shape, value (saturation) size, texture, orientation, and colour (Mackinlay, 1986). Animation is therefore highly complementary to techniques based around shape, colour and

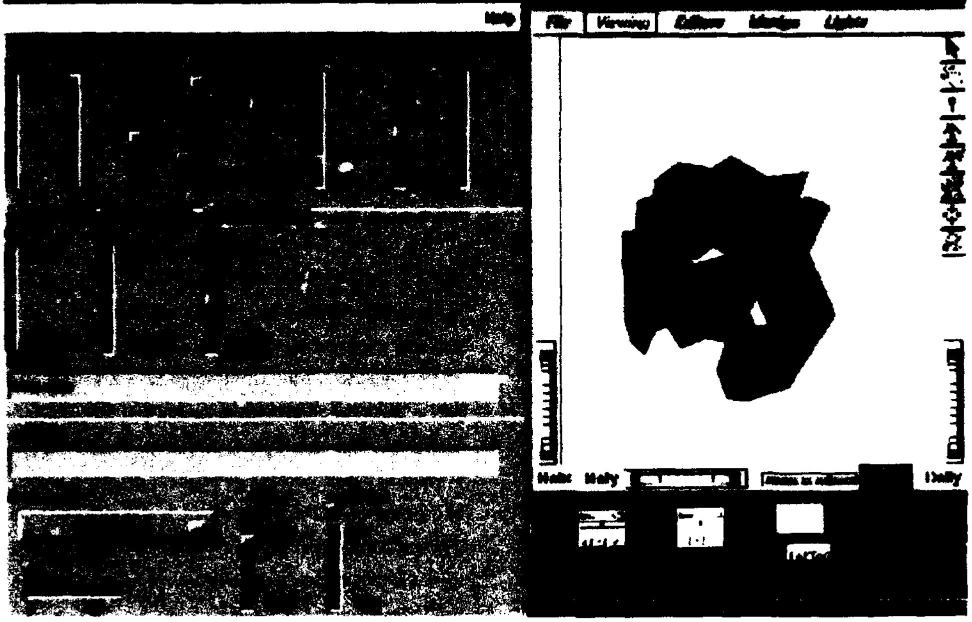


Figure 1 *The definition and control of an object interactor in the Explorer Environment. The shape describes the perimeter of a water body*

position. Rex & Risch (1996) describe a query language for the animation of geographic data.

2. *Styles of Interactors*

Two distinct methods for supporting interaction are described. Examples and discussion are given in Section 3.

2.1.1 Interactor Animation

The first method involves using a set of defined interactor tools (such as shown in Figure 1), which provide a visual link from one layer of data to another. The appearance of the interactor is animated in one of two possible ways, using either transparency or projection. In the former the interactor 'fades' in and out between two extremes (usually fully solid to fully transparent). The user is left with a 'visual imprint' but may also see all of the obscured data in a clear way at some point in the cycle. In the latter, the interactor is projected from one layer to another in small discrete steps, again with the aim of inducing 'pop out' as a link between data is established.

2.1.2 Layer Transposition Animation

The second method *animates the layers themselves*, so that one layer may be 'moved' through another by simple transposition along the Z axis. This moving layer actually *becomes* a complex interactor. At least one of the layers (although possibly both) must be in the form of a surface, otherwise they will not intersect gradually. It is the interaction of those variables used to provide the 'relief' that is most strongly emphasised. The careful assignment of attributes with which to construct and colour the surfaces is necessary for this technique to be effective.

2.2 Performance

Both methods are computationally intensive. In a fully rendered environment, as is provided by visualisation systems such as Iris Explorer, the scenes themselves can require substantial computing resources. Each visual parameter (*x, y, z, redness, greenness, blueness* and *transparency* as a minimum) use floating point precision. Coupled with the relatively large size of geographic datasets this makes con-

siderable demands on both the rendering speed and the amount of memory required. Although complex or large scenes may be rendered, any interaction with the scene will usually involve a good deal of 'swapping', and consequently will appear slow or 'steppy'. It becomes necessary to either restrict the extents of the data (windowing) or reduce the spatial resolution (sampling). On top of this considerable demand we now require smooth animation of a single object (the interactor). Where performance is inadequate, we must resort to constructing a video sequence off-line (a task that is easily automated) and then viewing this in real time.

3. *Results and Discussion*

Some example results demonstrating the use of both types of interaction are given below. It should be noted, however, that the printed page is woefully inadequate for representing an animated scene which is, by contrast, fully interactive (not to mention in colour); allowing the user to explore the data from any viewpoint, to animate and move objects and change their visual appearance.

An accompanying web site, http://www.cs.curtin.edu.au/gis/visualisation/geocomp.html, has been set up for this paper. Here, the figures used are viewable as high (and low) resolution colour images. The site also contains VRML scenes and video clips that may be downloaded and viewed from within Netscape in a more interactive manner.

3.1.1 Discussion of Interactor Animation

Animated interactors have proven useful for studying cause and effect relationships between different data layers to address questions such as: "*What evidence is there to support a particular (hypothesised) structure in the data?*" or conversely, "*How does a particular (known) structure appear in the data?*" Both of these questions refer to the *direction* of the interaction, being either *from* or *to* primary data. As a practical example, Figure 2 shows an interactor describing the extent of a salt scald (outbreak of surface salination) used between three data layers.

The lowest layer shows the source data from Landsat TM

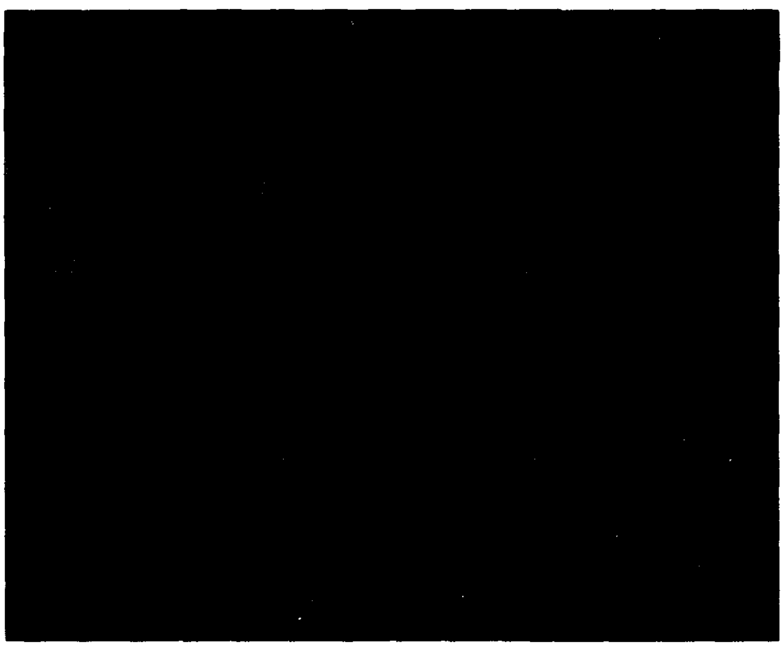


***Figure 2**: Three layers of data describing an agricultural region (see text for details). A semi-transparent interactor shows the location of a salt scald in all three layers.*

as a false colour image fragment, the middle layer is a landcover theme produced from the Landsat image by classification and the top layer contains some geographic objects known from ground truth. In this case, the salt scald represents a 'known' structure whose manifestation in the imagery is to be studied.

If the salt scald is *known* then the interactor shows how the scald is manifested in the thematic and image domain. If it is *hypothesised* then the interactor allows the user to visually study its appropriateness or plausibility.

Figure 3 shows a geological unit (the interactor) projected through a surface which represents magnetic anomalies. In this case, the geological unit is not *known* but is *hypothesised*. The aim is to study the plausibility of the unit by visually inspecting the scene to see if the magnetics surface contains evidence to support it. This type of task is often carried out by geo-physicists (but using more traditional techniques) in order to try to model the likely geological structure of a region.

Movie clips on the web site provide three animated examples using the same data as Figure 2. Each example shows a different type of animation of an interactor which represents the salt scald, using (i) transparency, (ii) projection and (iii) transparency and projection together.

Interactors have two major disadvantages. Firstly, they add visual clutter to the scene, causing a good deal of obscuration if not used carefully. Secondly, their geometry must be 'defined' beforehand.

The result of showing all interactors at once is to obscure *all* of the underlying data. To counter the cluttering, a mechanism is needed to establish the current focus of attention in the scene, so that only certain interactors are

***Figure 3**: An interactor describing a hypothesised geological unit is projected onto a magnetics surface.*

shown, and others are removed. Obviously, this mechanism must react to changes in the *focus of attention*. The specification for VRML-2 provides a mechanism whereby focus of attention can be supported interactively, by means of 'hot objects'; these can generate messages according to their proximity to the pointing device within the scene. It is possible to project interactors only from objects that are currently 'hot', as Figure 3 shows.

Interactors must be defined in terms of their perimeter, by a simple geometry, and it is this geometry that is then 'projected' into the Z domain. In many cases, the interactor represents a 'known' structure, such as a paddock boundary. In this case, its structure is explicit or implicit in the data (depending on whether a vector or a raster data structure is employed). The interactors used in the VRML example on the web site are automatically derived from a geological dataset in (Idrisi) image format.

3.1.2 Discussion of Layer Transposition Animation

Figure 4 shows instead the animation of the layers themselves, and requires a detailed explanation. Two surfaces have been constructed, the upper surface is a Digital Elevation Model (DEM) on top of which three channels of Landsat TM data have been draped (using a false colour assignment). The lower surface is more artificial and represents a wetness - greenness composite. Vertical offset is given by Landsat TM band four, and colour is provided from a surface water accumulation method applied to the DEM, coloured on a scale of green to blue, where blue represents the highest values (the "wettest"). The image in Figure 4 shows a single frame from the sequence as the lower surface is passed through the upper surface (the full animation may be downloaded from the web site). To the knowledgeable user, this can give information regarding likely environmental niches by showing the structural relationship between spectral response, water accumulation, height, aspect and slope, all of which are discernible (by differing extents) from the one animation. Particularly, the relationships between landscape structure and spectral response is emphasised.

This example raises a related issue: once the capacity to generate complex and artificial scenes is provided, it may be difficult for users to orient themselves within an unfamiliar 'space'. Experience to date has shown that some types of user may find this difficult, whilst others find it quite intuitive. It is a major premise of this style of interaction that the human visual system can 'orthogonalise' the various channels that together make up the stimulus space (Senay & Ignatious, 1994; Rheingans & Landreth, 1995). Put more simply, humans can learn to differentiate between the way variables are assigned to visual attributes.

Animating entire layers can cause much data to be obscured, so choosing the 'best' variables to assign to the Z dimension is paramount. This will obviously depend on the task at hand, so demands guidance in the form of either a high degree of user interaction or an expert system (Gahegan & O'Brien, 1997). Transparency may be used so that 'obscured' data is partially visible, but colour and shape information become weaker as a result and are consequently more difficult for the user to evaluate.

As final examples, Figure 5 shows an isolated magnetic anomaly (a geo-physical data source) for which magnetic intensity is doubly-encoded using height and colour, an example of *redundant assignment* to add emphasis (Bertin, 1981). The magnetics layer is passed through a SPOT panchromatic image of the surface to allow a study of the relationship between the anomaly and surface appearance. In this example, transparency is also used (on the SPOT data) so that the magnetics data is always partly visible. However, the transparency weakens perception of the surface structure.

4. *Conclusions, Ongoing and Further Work*

Simple visual paradigms, such as the use of animated interactors, are easy to understand and to communicate. The more complex paradigms available when animating entire layers are less so, but are capable of providing insight into highly complex relationships.

In order to visually analyse complex relationships, some

Figure 4 *Two intersecting environmental surfaces (see text for details)*

Figure 5 *A single frame from an animated magnetic anomaly surface as it is moved through a SPOT image*

form of visual encoding must be used, to give the required extra dimensionality to the stimulus space. This minimises the number of layers of data required, which in turn aids cognition. The assignment of data to the various layers and graphical objects remains a difficult problem which is yet to be fully solved and the use of expert systems, to aid the user and to cost out the huge numbers of alternatives, is currently being investigated (Gahegan & O'Brien, 1997). The software described here, in the form of Explorer scene graphs, modules and VRML scripts can be made available to other researchers on request.

Various schemes to utilise the remaining visual attributes of an interactor are possible. As examples, colour may be used to specify the direction of interaction, and opacity may be used to show the degree of confidence or uncertainty in a particular object (Howard & MacEachren, 1996).

Ongoing work also includes the design of various scenarios for exploring data, specifically addressing a user's needs when focussed on a particular object, group of objects or a dividing boundary between objects. Future developments will involve adding some adaptive behaviour into the visualisation environment, so that the needs of the user are catered for without lengthy or verbose system interaction.

5. Acknowledgements

The author would like to acknowledge the following research students at Curtin University: Vilya Harvey, Scott, Pleiter and David O'Brien, who have contributed some of the scenes used in this paper and included at the accompanying web site.

6. References

Bertin, J. (1981). *Graphics and Graphic Information Processing.* De Gruyter, Berlin, Germany.

Csinger, A. (1992). The psychology of visualisation. *Technical Report Series*, Department of Computer Science, University of British Columbia, Canada.

Duclos, A. M. and Grave, M. (1993). Reference models and formal specification for scientific visualisation. In: Palamidese, P. (Ed) *Scientific Visualisation: Advanced Software Techniques*, pp. 3-14.

Freidell, M. Kohhar, S., Lapolla, M. and Marks, J. (1992). Integrated software, process, algorithm and application visualization. *Visualization and Computer Animation*, Vol. 3, pp. 201-218.

Gahegan, M. N. (1996). Visualisation strategies for exploratory spatial analysis. Proc: *Third International Conference on GIS and Environmental Modelling*, NCGIA, Santa Barbara, USA.

Gahegan, M. N. and O'Brien, D. L. (1997). A strategy and architecture for the visualisation of complex geographical datasets. *International Journal of Pattern Recognition and Artificial Intelligence*, Vol. 11, No. 2, pp. 239-261.

Goodchild, M. F. (1992). Geographical data modeling. *Computers and Geosciences*, Vol. 18, No. 4, pp. 401-408.

Hartmann, J. L. (1992). Visualisation techniques in GIS. Proc: *EGIS '92*, Vol. 1, European Geographic Information Systems Foundation, Utrecht, Netherlands, pp. 406-412.

Haslett, J., Bradley, R., Craig, P., Unwin, A. and Wills, G. (1991). Dynamic graphics for exploring spatial data with application to locating global and local anomalies. *The American Statistician*, Vol. 45, No. 3, pp. 234-242.

Hearnshaw, H. M and Unwin, D. (Eds.) (1994). *Visualization in Geographical Information Systems*, John Wiley & Sons, Chichester, England.

Howard, D. and MacEachren, A. M. (1996). Interface design for geographic visualisation: tools for representing reliability. *Cartography and Geographic Information Systems*, Vol. 23, No. 2, pp 59-77.

ISO/IEC DIS 14772 (1997). The Virtual Reality Modeling Language: (VRML97). URL: http://vag.vrml.org/VRML97/DIS/index.html.

Jung, V. (1996). Fuzzy effectiveness evaluation for intelligent user interfaces to GIS visualisation. Proc. *First International Conference on GeoComputation*, Dept. of Geography, University of Leeds, UK.

Keller, P. R. and Keller, M. M. (1993). *Visual Cues: Practical Data Visualization*, IEEE Press, Los Alimatos, CA, USA.

Livingstone, M. and Hubel, D. (1988). Segregation of form, colour, movement and depth: anatomy, physiology and perception. Science, Vol. 240, pp. 740-749.

Mackinlay, J. (1986). Automating the design of graphical representations of relational information. *ACM Transactions on Graphics*, Vol. 5, No. 2, pp. 110-141.

O'Brien, D. L., Gahegan, M. N. and West, G. A. W. (1995). Information overload, the visualisation of multiple spatial datasets for mineral exploration. Proc. *GISRUK-95*, University of Newcastle, England.

Rex, B. and Risch, J. (1996). Animation query language for the visualisation of temporal data. Proc: *Third International Conference on GIS and Environmental Modelling*, NCGIA, Santa Barbara, USA.

Rheingans, P. and Landreth, C. (1995). Perceptual principles for effective visualisations. In: *Perceptual Issues in Visualisation*, (Eds. Grinstein, G. and Levkowitz, H.), Springer-Verlag, Berlin, pp. 59-69.

Robertson, P. K. (1990). A methodology for scientific data visualisation: choosing representations based on a natural scene paradigm. Proc: *IEEE Conference on Visualisation '90*.

Senay, H. and Ignatius, E. (1991). Compositional analysis and synthesis of scientific data visualisation techniques. In: *Scientific Visualisation of Physical Phenomena* (Ed. Patrikalakis), Springer-Verlag, Hong Kong.

Senay, H. and Ignatius, E. (1994). A knowledge-Based system for visualisation design. *IEEE Computer Graphics and Applications*, pp. 36-47.

Tang, Q. (1992). A personal visualisation system for visual analysis of area-based spatial data, Proc: *GIS/LIS' 92*, Vol. 2, American Society for Photogrammetry and Remote Sensing, Bethesda, Maryland, USA, pp. 767-776.

Upson, C. (1991). Volumetric visualization techniques. In: *State of the Art in Computer Graphics* (Eds. Rogers, D. and Earnshaw, R.), Springer-Verlag, pp. 313-350.

Interactive Exploration of Spatially Distributed Near Infrared Reflectance Data

Hayden Munro [1,2]
[1] Department of Information Science, University of Otago, P.O.Box 56, Dunedin, NZ.

Kevin Novins
[2] Department of Computer Science, University of Otago, P.O.Box 56, Dunedin, NZ.

George Benwell
Department of Information Science, University of Otago, P.O.Box 56, Dunedin, NZ.

Alistair Mowat
HortResearch, Ruakura Research Centre, Private Bag 3123, Hamilton, NZ.

Presented at the second annual conference of GeoComputation '97 & SIRC '97, University of Otago, New Zealand, 26-29 August 1997

Abstract

The display of a vector of data values at a set of spatially-distributed sample points presents some interesting visualisation problems. Typical display devices provide only two spatial di mensions plus colour, making it necessary to design new methods for representing the data. This paper describes a tool under development that allows users to visualise the spatial ripening characteristics of fruit. Sugar, acid and moisture content can be measured using non-destructive Near Infrared Reflectance (NIR) analysis techniques. We introduce the notion of *spectrum* and *spatial* tools and show how they may be combined to form a flexible visualisation environment for exploring NIR data. These notions may be generalised to areas such as LANDSAT imagery. We expect that high performance computing systems will enable us to extend our tools so that they can operate at a wide range of spatial scales.

1. Introduction

Scientific visualisation can be described in two ways: as a tool for discovering and under standing, and as a tool for communicating and teaching (DeFanti, 1990). It is used to present information to users in visual forms that appeal to their intuitive understanding. Thus, visualisation tools facilitate the extraction of knowledge from complex datasets. This paper describes a tool that is being developed to allow users to visualise data about the ripening characteristics of fruit. These characteristics, such as sugar, acid and moisture content can be measured using non-destructive Near Infrared Reflectance (NIR) analysis techniques (Ciurczak, 1995, Murakami, 1993, Murakami, 1992). By comparing the information gained by NIR analysis with the physio chemical properties of the fruit it is possible to relate specific NIR spectral features to desirable product attributes. Once this has been done NIR analysis will be able to be used as an objec tive measure of fruit quality. This information could then be used to manage fruit development and storage processes to maximize market acceptance. It may also be possible to predict fruit maturity and post-harvest characteristics prior to harvest.

The multidimensional nature of the NIR data introduces some interesting visualisation prob lems. The data is four dimensional, whereas the display device only provides two dimensions. It is therefore necessary to discover two dimensional methods for representing the data (McCormick, 1987). Also, the users of the application wish to explore the dataset to discover new features, so it is important to create an application with a high degree of interaction to assist them. In order to tackle these problems this implementation provides a suite of interactive visualisation tools rather than a single display mode. In addition users are given the power to modify the display for their own purposes.

The approaches described here are applicable to all visualisation systems that work with a set of scalar values measured at spatially-distributed sample points. For example, the reflectance data gathered from a fruit using an

NIR probe is exactly analagous to spectral reflectance data measured by a satellite orbiting the earth.

The current tool is useful for exploring spatial relationships in the NIR data across individual fruit. Ideally, we would like to analyse spatial relationships at a wide range of scales, for example comparing different fruit from the same tree, different trees in the same orchard, or different orchards in the same geographic region. Data exploration of this kind is not practical using our current workstation-based system. It will require the high RAM capacity and rapid paging abilities of high performance computing systems. The work described here is part of a continuing joint research between the University of Otago and the Horticulture and Food Research Institute of New Zealand.

2. Data collection

The data for this project is provided by HortResearch. It consists of diffuse reflectance spectra, comprised of individual light intensities measured at 0.5nm intervals, collected from specific locations on the sub-surface of intact fruit. The spatial coordinates of these locations were also provided. So far, datasets have been collected from Gala apples and kiwifruit.

The spatial coordinates were obtained using a Polhemus FasTrak T M device (Smith, 1995). Each fruit was placed inside a cylinder with the major axis perpendicular to the cylinder's straight edge, as shown in Figure 1. A longitudinal line of nine holes had been drilled at 15 degree intervals around the cylinder's circumference. The stylus of the tracker is pushed through each of the holes to measure a position on the fruit surface. These point locations correspond to one eighth of the fruit. The fruit is then rotated 45 degrees to collect the next set of points, and the process is repeated until the whole fruit surface has been sampled. In total 72 points, distributed across eight points of latitude and nine points of longitude, are selected for spectral analysis. This collection method results in a 3D dataset in cylindrical coordinates. These must be converted to Cartesian coordinates before visualisation.

After the FasTrak(TM) device is removed from the guide holes, a Near Infrared Reflectance (NIR) probe is inserted to record the diffuse visible-to-NIR spectrum at the fruit's subsurface at each of the 72 points. The diffuse visible-to-NIR spectrum was sampled at 1100 intervals in the spectral range of 508nm to 1026nm using an Ocean Optics miniature fibre optic probe and charged couple device (CCD) spectrophotometer. The sampled spectrum shows the amount of energy that is reflected at each different measured wavelength (Hall, 1988). There is a minor error term between intensity measurements within a spectral reading. This error is related to the noise characteristics of the 1100 elements contained in the CCD array sensor of the spectrometer.

Commercially available spectral analysis packages such as Grams/32 (Glactic Industries Corp., Salem, NH, USA) may be used to compare reflectance spectra for different points on the fruit. These tools are not sufficient for our task because they do not preserve the spatial relationships between the points that are being compared.

3. Visualisation methods

The data described in the previous section can be considered four dimensional: the points on the fruit's subsurface where wavelengths were measured must be described in 3D in order to maintain their spatial relationships. The intensities measured over the sampled spectrum are the fourth dimension.

Since the data has more dimensions than the display, no single visualisation method will be able to display all aspects of the dataset simultaneously. For this application, the user cannot determine in advance which view will be best. Fortunately the computer's flexibility allows the development of a suite of visualisation tools that present different conceptual views of the data. By using the tools interactively and in concert the user can explore and discover features of interest.

The tools currently provided are divided into two categories: Spectrum tools and spatial tools. Spectrum tools display the intensity at all measured wavelengths for a single point on the fruit surface. Spatial tools display 3D fruit

Figure 1. Selecting spatial points on a fruit using a Polhemus tracker

geometry and intensity values for a single wavelength at all surface points. These categories are described in detail in the following subsections.

The tools were developed using the OpenGL library and the Tcl/Tk scripting language. OpenGL is a standard graphics library designed for real-time applications (Neider, 1993). It is designed to encourage the development of portable programs. OpenGL is a library of functions that can be called from a high-level language, in this instance C is used. OpenGL makes it particularly easy to create and interact with 3D polygonal models.

Tcl/Tk was also chosen for its flexibility. Tk provides a package of tools that are used for creating application interfaces. Tcl is an interpreter that builds the program's interface from command scripts (Ousterhout, 1994, Welch, 1995). Using scripts rather than library calls allows the interface to be modified while the application is running. This feature creates a small decrease in terms of speed, however this is not significant in our application.

The software tools described here are available for many platforms; for the implementations described here, a Silicon Graphics Indigo workstation was used.

3.1 Spectrum tools

Spectrum tools display the intensity at all wavelengths for a single point on the fruit surface. This is essentially a 2D graphing problem: For each of 1100 wavelength samples, we have a single intensity value. We can present the data as a 2D scatterplot, a line drawing, or a histogram. These are presented in Figure 3. Because of its high contrast, we have found the spectral histogram tool to be the most useful.

3.2 Spatial tools

The following views have a common trait that sets them aside from the spectrum tools: Each spatial tool shows multiple points on the fruit surface, but intensity at only one wavelength. This enables comparison of intensities between points, and provides an opportunity to observe their spatial relationships.

3.2.1. 3D Scatterplot View.

The spatial coordinate data can be plotted in a virtual 3D space, as shown in Figure 4. A perspective projection transforms the 3D model into a 2D picture for display (Foley, 1990). The user can interact with the scene using a virtual trackball (Glassner, 1990). The mouse manipulates the viewing position so the model can be seen from any angle, giving the impression of a three dimensional object.

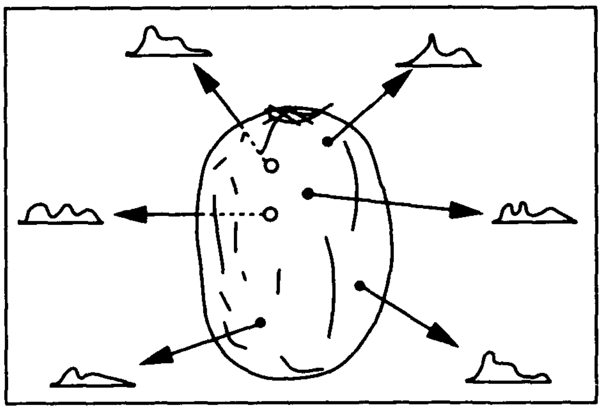
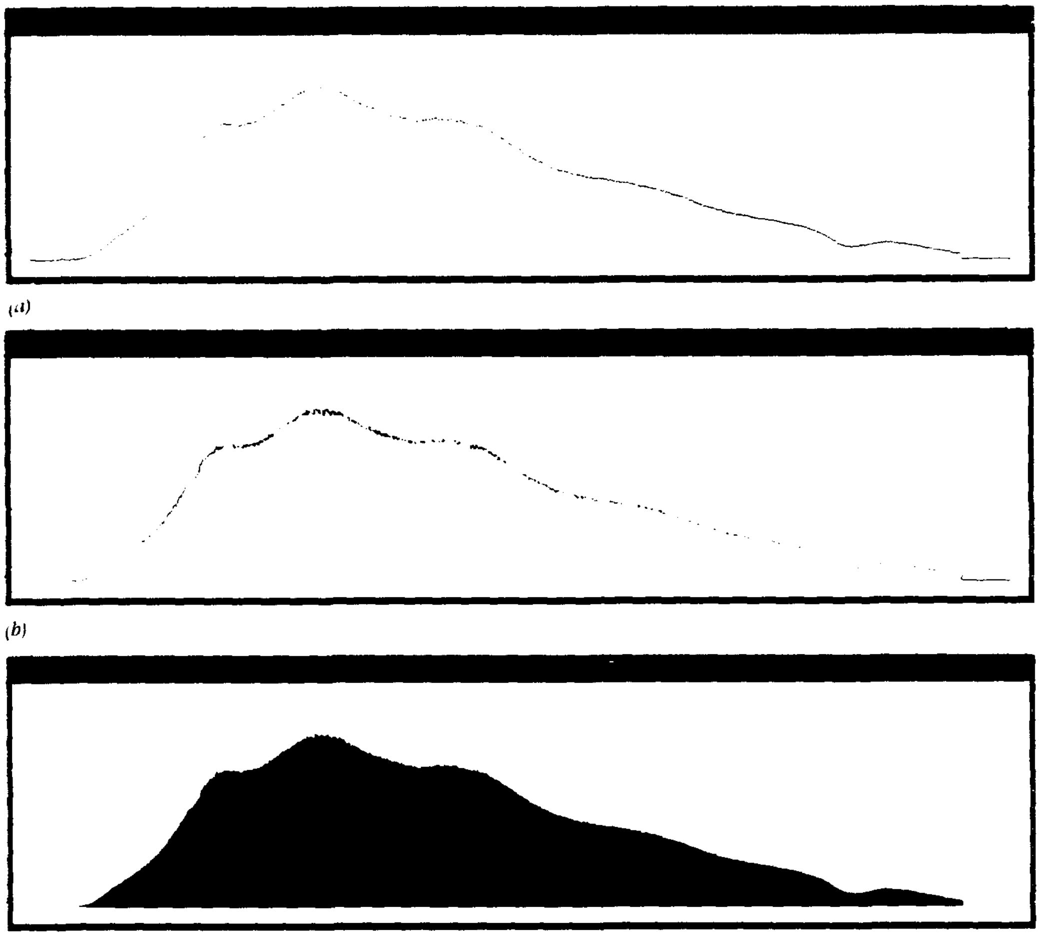


Figure 2. Data is gathered across the NIR spectrum for each sample point, so wavelength becomes a 4th dimension to visualise

Plotting the points in 3D is not sufficient; it is necessary to find a way to show the spectral data. This can be achieved by displaying the intensities at a single wavelength as colour information. By mapping the largest intensity



(a)

(b)

(c)

Figure 3 The spectrum tools (a) the point plot, (b) the line plot, and (c) the spectral histogram

to white and the smallest to black, these and the intermediate values can be shown as levels of gray on the fruit model. A slider can be used to select the individual wavelength to display.

Representing intensity with grayscale has a drawback. If the model is shown on a 24-bit monitor, there is a restriction of 256 grey values. This means that it is only possible to distinguish between 256 intensity values in the dataset. By using the red, green, and blue channels independently it is possible to increase this number by a factor of six. Ware (Ware, 1988) suggests that the lack of colour resolution is a minor problem when compared with systematic errors that can arise from interpretive effects of the human visual system, such as simultaneous contrast. With this in mind it is more important to reduce these perceptual effects than it is to reduce quantisation of the displayed data. In accordance with this philosophy an approximation to the visual spectrum is made available as a colour map for the wavelength data. The 3D scatterplot view with the visual spectrum colour map is shown in Figures 5a and 5b.

3.2.2. 3D Model View.

The points chosen for each fruit represent a rather sparse

sample. In order for the fruit to be displayed as a "skin" the intensity information is interpolated between points. The method used to interpolate this colour information is known as Gouraud shading (Foley, 1990). Gouraud shading is a popular technique because it is simple and can be computed using graphics hardware. It is useful to extend these points to form a skin for two reasons. First, the shape of the model is clearer if it is represented as a skin. Second, the point colour is interpolated across an area, making it easier to see. The skin is made from polygonal surfaces constructed using adjacency information extracted from the latitudes and longitudes of each point. The 3D scatterplot view is compared to the 3D model view in Figure 5.

3.2.3. The Map View.

The 3D tools are useful for viewing the fruit in a virtual environment but it is restrictive since at any time the back part of the fruit is occluded, leaving the data only partially visible. It is possible to "unwrap" the model view so that the full surface of the fruit is shown in one picture. This is shown in the map view. An example map view is shown in Figure 6. The data collection method provides a natural latitude-longitude coordinate system for the data. In principle, this partitioning allows us to use any of the techniques developed by cartographers for mapping the earth. In order to "flatten" the fruit surface our current tool simply plots longitude as the x-axis and latitude as the y-axis in what is known as an equirectangular projection (Snyder, 1993). All 2D maps of curved objects will contain distortions, and while accuracy of the equirectangular plot is high at the fruit's equator it decreases rapidly as the poles are approached. In this application the effects of distortion are reduced somewhat since the poles of the fruit are not sampled.

3.2.4. The Height Field View.

The map view reduces visual complexity and allows us to use the 3D graphics hardware for another purpose. In the height field view intensities are interpreted as altitudes, and the data is displayed as a 3D relief map. This artificial terrain helps to clarify the relative distance between in-

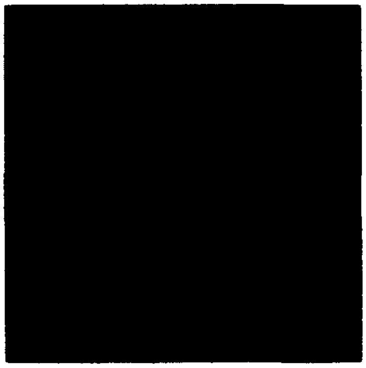


Figure 4 The 3D spatial points

tensity values, which is difficult to determine when only colour information is used. The height field can be viewed from any angle by manipulating a virtual trackball. The heights also respond interactively to movement of the wavelength slider. It is important to incorporate a stationary reference for this view; otherwise the surface appears as if it is floating in space. The stationary reference is created by vertical lines below the height field in Figure 7.

A new slider is introduced for the height field view. It controls the height of the maximum value that is in the dataset, acting as a scaling constant for the virtual altitudes. This is provided so the user can exaggerate the disparity between similar intensity values that are displayed. It is important to note that applying heights to the intensity values in the 3D model view would not be beneficial. The 3D object is too complex to act as an adequate reference for the changing amplitudes. It would be difficult to distinguish between bumps that represent fruit geometry and bumps that represent intensity data.

4. View integration

The views described in the previous section are combined to create a single application. Each view presents different aspects of the data, so they are made to occupy separate windows that are simultaneously visible. To support exploration using multiple windows, the parameters from each view are linked.

Each point shown on the map view can be selected by clicking or dragging the mouse in the display window. The

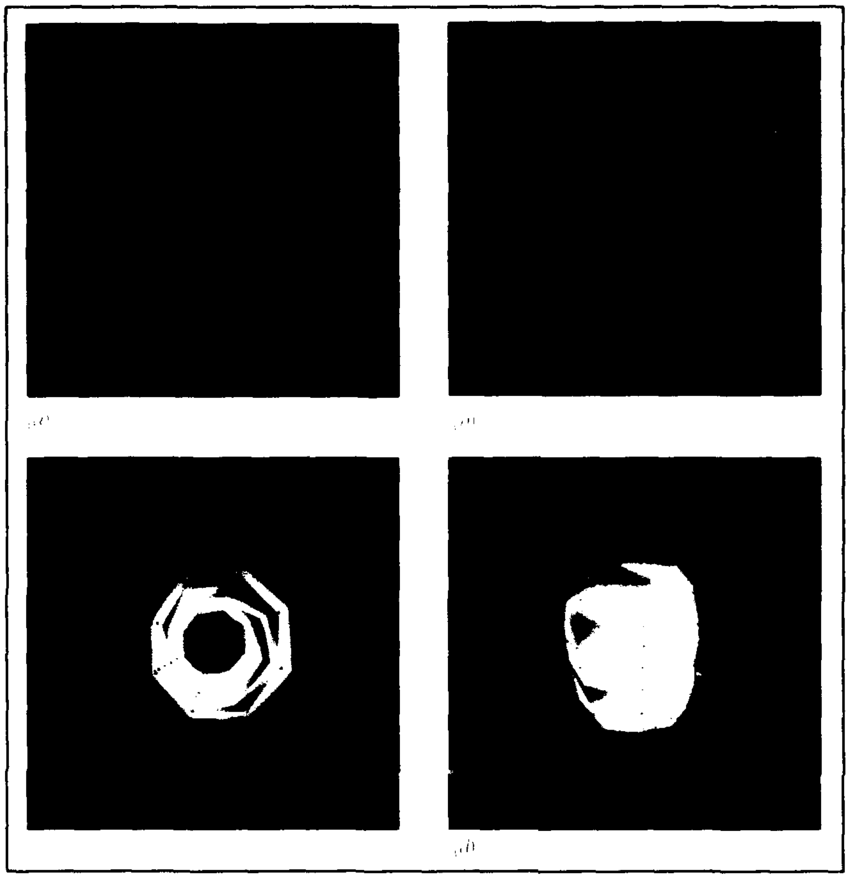


Figure ... Comparing 3D scatterplot and 3D model views. 3D Scatterplot views for the 7??nm response of a ... are shown from two different angles in (a) and (b). In (c) and (d) the corresponding 3D model views are shown.

closest point to the mouse is highlighted and the spectrum data relating to that point is used to create a spectral histogram. As the mouse is dragged between points the histogram changes interactively. It is possible to select and drag a wavelength indicator on the histogram. As this is done, the wavelength shown in the map view is updated accordingly. This single environment, illustrated in Figure 8, is maintained to avoid confusing the user.

In the current application, the 3D model view and the height field view are also linked to the wavelength selected in the spectral histogram view. A screen shot of the application along with descriptions of the interaction methods is given in Figure 9. The cohesion between the views provides a fast and effective way to view the data.

5. Conclusion and future work

The integrated application was recently presented to a group of scientists at HortResearch. Their response illustrated that the views create a visualisation environment that encourages exploration of the dataset. Interacting with the raw data graphically gives the user a rapid understanding of the nature of the data. The provision of multiple views has created a great level of flexibility for exploring the dataset. The presentation stimulated many ideas about

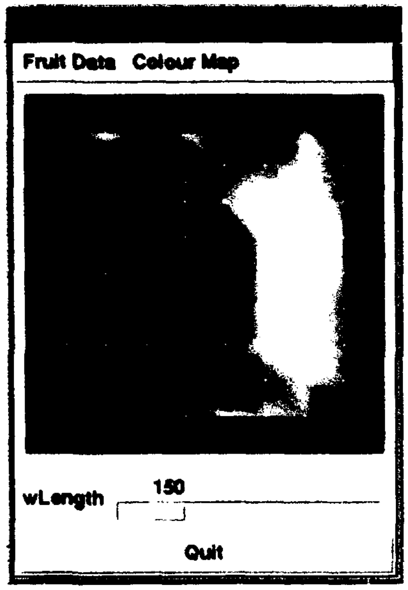


Figure 6 The map view

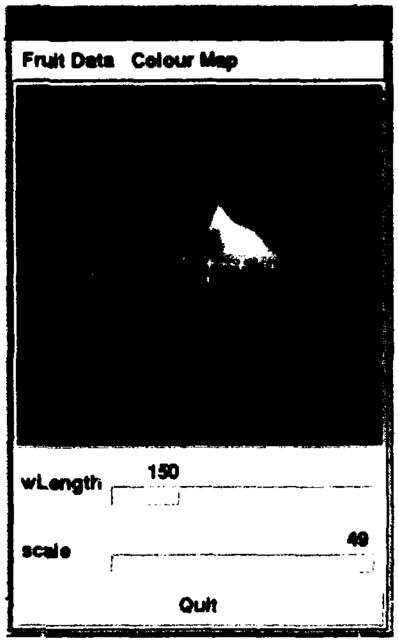


Figure 7 The height field view

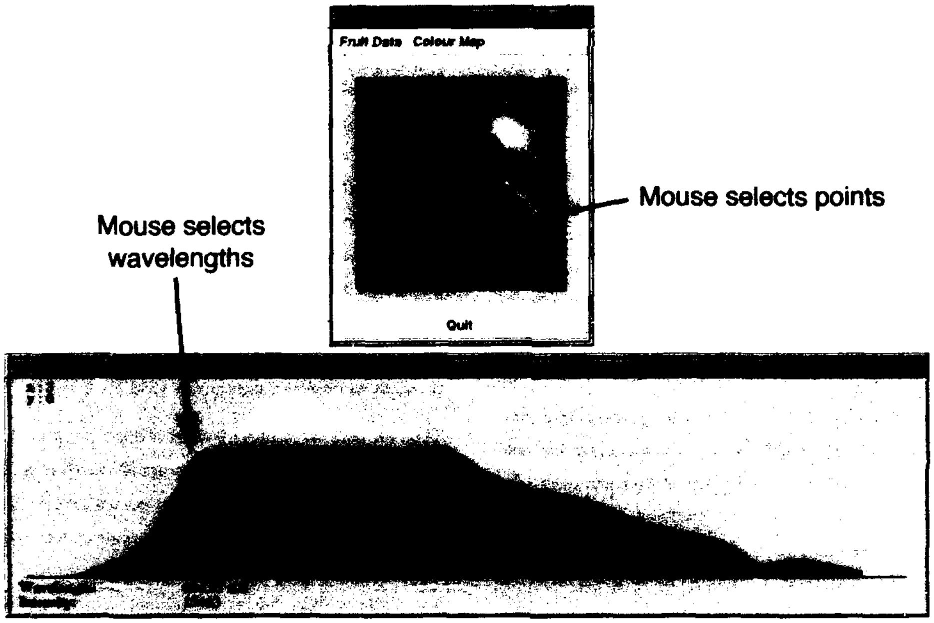


Figure 8 The interaction methods for map and spectral histogram views

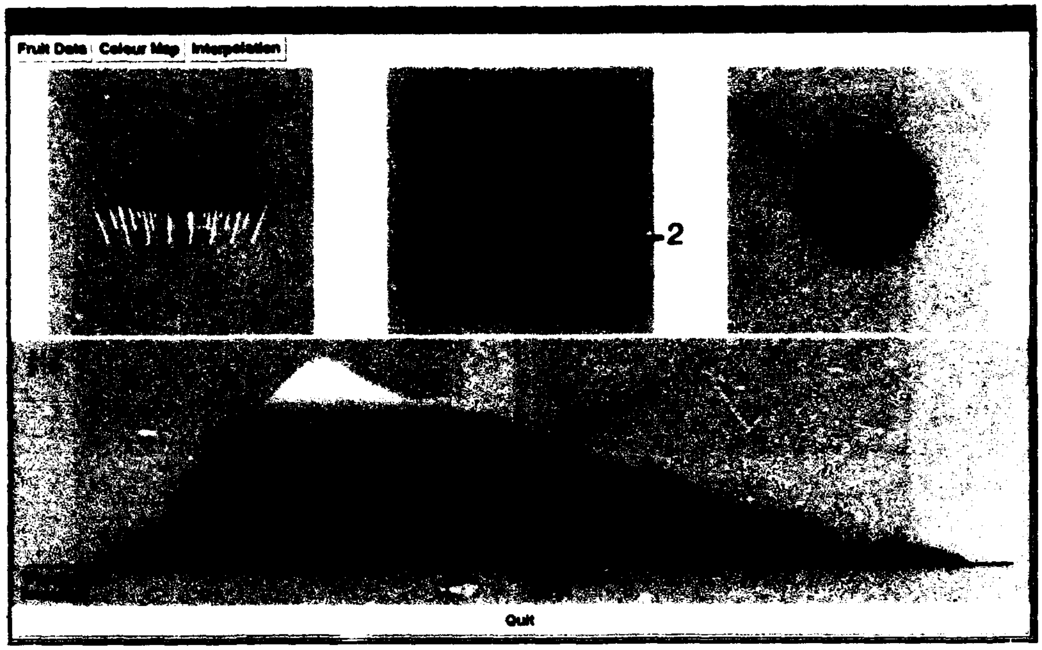


Figure 9. The integrated application with four interaction methods. In the height field view (1) and the 3D model view (3), the mouse acts as a virtual trackball to rotate the object. Fruit sample points are selected by the mouse in the map view (2), and the selected wavelength is changed by dragging the indicator in the spectral histogram view (4)

how the tools can be enhanced and has defined a clear path for their future development. The tools are considered promising and the project is continuing.

We would like to tune some of the interaction paradigms that are used in the system. For example, it would be useful to be able to select sample points in each of the spatial displays, not just the map view. It would also be helpful if the spatial tools automatically centered the currently selected sample point. This could be done by scrolling the map and height field views and rotating the model view.

Our experience indicates that users gain accurate information about NIR datasets using the visualisation tool. Controlled perceptual studies are necessary to ensure that the user's impres sion of the visualised data matches the reality of the raw data. Possible areas of inaccuracy include the use of a quantised colour scale to display intensity, linear interpolation of intensity between sample points, and the equirectangular map projection.

Because NIR data collection is inexpensive and non-invasive, it is possible to measure the response of an individual fruit throughout its maturation process without removing it from the vine. Portable scanners are already under development (Martinsen, 1996). The most exciting follow up re search will involve the display of NIR data at varying spatial scales and across the temporal dimension. Analysis of this type of data may enable researchers to predict post-harvest quality prior to harvest. They may also be able to provide farmers with advice on how to situate trees and orchards to maximise high-quality yield.

The NIR data for a single fruit at a single point in time requires about 317 kilobytes of storage. When this number is multiplied by the number of apples in a tree, the number of trees in an orchard, the number of orchards in a region, and the number of days to maturity, the memory requirements quickly reach the terrabyte range. In order explore such a dataset interactively, we will need very large RAM

spaces, fast paging capabilities, and high-speed graphic hardware capable of implementing a zoomable interface (Perlin, 1993).

6. Acknowledgements

The software products used in this implementation are Tcl/Tk and OpenGL (a trademark of Silicon Graphics Inc.).

References

E.W. Ciurczak (1995). Use of near-infrared spectroscopy in cereal products. Food testing and analysis, 1(2).

T.A. DeFanti, M. D. Brown, and B. H. McCormick (1990). Visualisation, expanding scientific and en gineering research opportunities. In G. M. Nielson, B. Shriver, and L. J. Rosenblum, editors, Visualisation in Scientific Computing, pages 32-47. IEEE Computer Society Press.

J. D. Foley, A. van Dam, S. K. Feiner, and J. F. Hughes (1990). Computer Graphics, Principles and Practice. Addison Wesley, 2nd edition.

A. S. Glassner (1990). Graphics Gems, chapter 8, pages 462-463. Academic Press.

R. Hall (1988). Illumination and Color in Computer Generated Imagery. Springer-Verlag.

P. Martinsen and P. Schaare (1996). A near infrared imaging spectrometer. Image and Vision Computing New Zealand, pages 109-114, August.

B. H. McCormick (1987). Computer graphics: Special issue on visualisation in scientific computing. Technical Report 21(6), ACM SIGGRAPH, November.

M. Murakami, J. Himoto, and K. Itoh (1993). Analysis of apple quality by near infrared reflectance spectroscopy. Journal of faculty of agriculture, hokkaido university. September.

M. Murakami, J. Himoto, M. Natsuga, and K. Itoh (1992). Analysis of pumpkin quality by near infrared reflectance spectroscopy. Journal of faculty of agriculture, hokkaido university, 65(4):359-366.

J. Neider, T. Davis, and M. Woo (1993). OpenGL Programming Guide. Addison Wesley, 1993.

J. K.. Ousterhout (1994). Tcl and the Tk Toolkit. Addison Wesley.

K. Perlin and D. Fox (1993). Pad: An alternative approach to the computer interface. In Computer Graphics Proceedings, Annual Conference Series, pages 57-73.

G. S. Smith and J. P. Curtis (1995). A fast and effective method of measuring tree structure in three dimensions. Technical report, The Horticulture and Food Research Institute of New Zealand Ltd.

J. P. Snyder (1993). Flattening the Earth. University of Chicago Press.

C. Ware (1988). Color sequences for univariate maps: Theory, experiments, and principles. IEEE Computer Graphics and Applications, 8(5):41-49, September.

B. Welch (1995). Practical programming in tcl and tk. Draft Manuscript.

Rule-Based Modelling In GIS

Dr. David Pullar
Geographical Sciences and Planning
The University of Queensland. Qld. 4072
AUSTRALIA
+61 7 33656522
D.Pullar@mailbox.uq.edu.au

Presented at the second annual conference of GeoComputation '97 & SIRC '97, University of Otago, New Zealand, 26-29 August 1997

Abstract

Geographic Information Systems (GIS) provide functionality for visualising, managing and manipulating spatially referenced data. For straight forward spatial queries users perform these functions through a graphical user interface, for more advanced spatial models the user is often required to write a program as a formulation of the problem. The objective of this paper is to find out what type of programming language is suitable for users who do not have extensive programming experience, and yet provides powerful modelling capability. The paper reviews different types of programming paradigms and the abstractions they support. It argues that a combination of decision rules and object-oriented paradigms offer a clear and effective language interface to GIS. The rules are structured and presented in a tabular form to simplify their specification. The approach emphasises a style of programming that is attuned to spatial data models and programs are easy to comprehend.

1. Introduction

A necessary part of solving problems with computers is to express them in a formal way. The appropriate computer tool to solve geographical problems is a Geographic Information System (GIS). Such systems provide the basic functionality for visualising, managing and manipulating spatially referenced data. Problem solving is expressed using a computer language either provided by the system or one that interoperates with the system. Users of GIS are faced with the task of writing programs as a concrete formulation of their particular problem for not only advanced spatial analysis problems, but also for many *ah hoc* queries. Given that most users of GIS are professionals in a land related discipline, and not professional programmers, then it is important to make the language interface to a GIS as easy to use and intuitive as possible.

Novice users often find writing programs to be a daunting task. Assuming there is a systematic or scientific approach to the problem being solved, expressing this in a program is still a difficult intellectual activity. This paper has identified two fundamental reasons for this:

1. *Representation mismatch* between the object level representation of spatial data in the programming language and the application view. This problem occurs when the application presents information in one way but the programming environment to access and manipulate that information is different. A popular way to present information in GIS is as a map organised into thematic layers, whereas in a programming environment the user is presented with tables containing records. This also affects the way queries are expressed. For instance the application interface provides associative access by querying feature properties, whereas the programming language may provide data access by retrieving records based upon relative record numbers in a table.
2. *Problem specification mismatch* between the way a user expresses a problem and how programming languages implement the solution. This occurs when a users expresses a problem in set theoretic terms, yet is forced to resolve the problem in an application program as a sequence of operations on individual records. Many

programming languages compel a programmatic approach to solve a problem involving detailed actions in strict operational order.

This paper examines programming-language integration with GIS. We consider elements of programming including control structures, data structures, arithmetic, and so forth. In particular we explore the potential of decision tables to express query and modelling problems in a conceptually intuitive way. Decision tables express condition-action clauses in a tabular form. Arentze *et al.* (1995) show this to be a flexible technique for decision support in facility planning. We further explore their integration with operational semantics of GIS. We propose that decision tables be used in combination with object-oriented methods to provide a very direct and concise way to solve geographical problems. To demonstrate the concept, a prototype development is described where decision tables are integrated with the programming framework used by a commercial GIS.

The outline of this paper is as follows. Sections 2 and 3 characterise programming languages by the type of information abstractions they support. It is concluded that a combination of decision rules and object-oriented access to spatial features provides a uniform and convenient notation. Section 4 advocates the use of decision tables as a presentation style. Examples for a non-trivial spatial query demonstrate the concepts.

2. Background

Programming paradigms can be characterised according to: i) the data abstractions, and ii) the procedural abstractions they support.

1. *Data abstraction* refers to the way information content is represented. Low level languages use simple value types, like integers and reals, while high level languages support more abstract object types and data modelling relationships. Objects types impose a class definition to describe structural properties (state information) and behavioural properties (operations). Data modelling relationships define how objects may be related. This includes associative relationships for structural linkages between objects, composition relationships where objects form part of an aggregation hierarchy, and generalisation relationships where objects share common semantics in an inheritance hierarchy.
2. *Procedural abstraction* refers to the way actions are defined and controlled within the programming language. Low level languages solve problems in terms of *imperative* commands. The program code describes exactly how to solve the problem as a rigidly controlled set of detailed actions. Program control is expressed by either sequential instructions, repetition or branching conditional constructs. Examples of low level languages include FORTRAN and C. High level languages solve problems in a *declarative* fashion. The program code specifies the desired outcome or goal. Program control is less rigid and may be based upon reasoning to prove a hypothesis. Examples of high level languages include C++ and PROLOG.

So what are the best language characteristics to use in GIS? We reduce the scope of this question by focussing on a programming paradigm that is designed for a typical GIS user, namely one who does not have a significant amount of training in programming techniques. Any procedural abstractions would need to be implicit to harmonise with the way a user attempts to solve a problem, and should exercise a reasonably obvious method of control over spatial features. It would support a range of queries on spatial databases without the user needing to understand the intricacies of computer algorithms. The data abstractions used within the language need to be tightly interweaved with the way information is managed and manipulated at the user level. In modern systems this means the language must harmonise with geographic data modelling methods and with user interface paradigms used to manipulate geographic information.

3. Programming Paradigms

Programming languages employ different types of data abstractions and procedural abstractions. Different types of languages stress one characteristic over another. Four main paradigms are identified:

- Logic Programming
- Functional Programming
- Rule-Based Programming
- Object-Oriented Programming

3.1 Logic Programming

Logic programming applies rules of exact logic to solve problems, or to be more exact it applies rules of first order predicate logic. Problems are expressed as statements to represent things that we believe about the world. The statements are composed of a set of logical terms and logical connectors. The rules for evaluating statements are given by a truth table shown in Figure 1.

In first order predicate logic all objects belong to a single universe. This leads to a characteristic of "flatness" in pure logical languages. All objects are universal and so are the axioms by which they are related. There is no procedural abstraction in first order predicate logic.

In practice, logic programming languages use some procedural mechanisms to interpret logical statements. The most popular of these programming languages is PROLOG (Bratko, 1990). A logical statement is expressed as a Horn clause consisting of a conclusion head "C" and several conditional terms "B" in the body. They have the form:

"B_1 and B_2 and B_3 ... and B_N implies C"

Different combinations of a head and body create three types of clauses: queries, rules and facts. The fundamental form of programming control is a query that is answered by searching for matching facts, or rules whose heads match the query and whose body may be proven. This ability to search through a set of facts and to further deduce relations from rules gives PROLOG its deductive capability.

Terms		*AND connector*	*OR Connector*	*Implication*
p	q	$p \wedge q$	$p \vee q$	$p \rightarrow q$
true	*true*	*true*	*true*	*true*
true	*false*	*false*	*true*	*false*
false	*true*	*false*	*true*	*false*
false	*false*	*false*	*false*	*true*

Figure 1: Truth Table.

The power of PROLOG-like languages to express both spatial queries and spatial models has been well demonstrated. LOBSTER is an early example of a prototype system that used PROLOG as the language interface to query a spatial DBMS (Egenhofer, 1990). The prototype provided a high level language to manipulate symbolic representations of spatial features. This was possible because the DBMS was able to handle complex record structures, and user defined functions could be programmed as builtins to the PROLOG interpreter. Spatial data types for points, lines, areas, and surfaces were defined in the DBMS and manipulated at a semantic level by the rules and facts expressed in Horn clauses. All low level access to spatial data and spatial manipulation is handled by the builtin functions. This ability to include declarative expressions of spatial queries within a logic language is viewed as a key requirement by other researchers (Abdelmoty *et al.*, 1993).

3.2 Functional Programming

Functional programming is based upon mathematical concepts of mapping functions. A function maps object values from one domain to another. This is expressed formally f:X®Y, the function f maps object values from the domain X to the domain Y. The object returned by a function depends only on its arguments. In addition functions do not induce any side effects so all state information evolves in an explicit and controlled way. This trait is known as *referential transparency*. Any transformations on objects are handled by explicitly returning new objects. This has a bearing on the data and procedural abstractions used by functional languages. Both rely upon mapping functions to express structural and behavioural relationships.

Advanced functional languages have a powerful expressive quality with the ability to use higher order functions (a function of a function of a .....) to perform symbolic manipulation and proofs in programs. Functions are also treated as first class objects so they may be used as arguments and may be the return value from a function. A mathematical style of programming is ob-

tained by using algebra expressions instead of function names. Examples of functional languages include LISP, this manipulates objects as a list of unitype symbols, and ML that is a language that manipulates objects with more advanced data types (Paulson, 1996).

A GIS database perceived and manipulated by a functional language is viewed as a collection of objects together with a collection of functions. This has not proven to be a very attractive quality for feature-based GIS applications as there is not sufficient selective distinction between the different operations permitted on various types of spatial features (ie. point, linear, and area features). However, GIS applications that use a simple image-based structure are more predisposed to this type of manipulation. Map algebra is an example of a function-oriented language used in GIS for manipulating and analysing surface data (Tomlin, 1991). Map algebra uses a set of conventions to provide finer interpretation of the geographic locations (ie. local, neighbourhood, zonal) but these are still manipulated by functional transformations. Map algebra has the advantage of a straight forward notation and is very useful for developing models of spatial interpretations.

3.3 Rule-Based Programming

Rule-based programming is a special case of logic programming. The language is based on a procedural scheme with the canonical condition-action form:

IF condition-pattern THEN actions.

The left-hand side consists of several conditions that return a logical result. The right-hand side consists of several actions. Actions can fire other rules, establish new facts, and perform procedural operations. Rules express relationships and meta-information. Rules are grouped in rule-sets known to the inference engine. The engine works in a continuous loop, at each cycle a rule that matches some condition-pattern is chosen and the related actions are fired. The execution stops when no more rules are fireable.

Rule-based programming uses a simple procedural abstraction to search for goals that satisfy the condition-pattern and then subsequently firing the action clauses. Queries are solved as proofs computed from the facts and rule set. Rule-based programming does not directly support data abstractions but relationships can be expressed by meta-rules.

Rule-based programming provides a model of the decision process that suits a range of problems used for spatial reasoning (Scarponcini *et al.*, 1995). The techniques have been used in several *ad hoc* system developments for decision support (Lowes and Bellamy, 1994) (Davis and McDonald, 1993).

3.4 Object-Oriented Programming

Object-oriented programming (OOP) is based on concepts for *objects*, *classes*, and the *inheritance* mechanism between classes. An object is an instance of a class to hold all related state information. Since objects can reference other objects, it is possible to build compositions of more complex objects. The classes in a program define categories of objects which share the same state information and procedural interfaces. Inheritance provides a relationship between classes based upon a taxonomy hierarchy. These organising principles are formally based upon classification theory.

OOP has become very popular as it provides a mental leverage for designers to encapsulate the structure and behaviour of design problems as objects. Data abstraction is supported through associative references to express structural relationships between objects, and class inheritance. Procedural abstractions are provided in two ways. The permissible actions on an object, and a configuration of objects, are integrated as part of the object class description. But the final implementation code still uses low level procedural mechanisms to perform operations in sequence, by conditional branching, or within an iteration. A disadvantage is that these control constructs involve the introduction of state variables to hold computational values between operations and procedures.

Writing a program in an OOP language does not necessarily make the program object-oriented. But in general programs incorporate object-oriented design principles

(Rumbaugh *et al.*, 1991). OOP is especially suited to problems where these is a large number of entities to be modelled, each with complex structural relationships and operational semantics. In recent years OOP has made a significant impact on graphical user interfaces (GUI's) and the application programming environment. Desktop GIS's often use object-oriented concepts in the user interface and application programming environment. But in most cases spatial data handling is still based upon a geo-relational model, and so data abstractions such as association and inheritance are not applied to the spatial data. Morehouse (1990) discusses the implications and difficulty of having true object-oriented modelling semantics for spatial databases. The OpenGIS Specification (OGC, 1997) incorporates object-oriented geo-processing concepts. The full development of models to allow user defined schemas will require information representation specified by data dictionaries, schematic catalogues, geometry rules, etc. This technology specification will have an impo[illegible]nt im[illegible] the adoption of object-oriented data abstractions within GIS programming languages.

3.5 Summary

Different programming paradigms may be characterised by the data and procedural abstractions employed. The four paradigms and the types of abstractions supported are summarised in Figure 2. Note that most language implementations use a combination of programming paradigms, or programmers adopt a style suited to one or another programming paradigm. Therefore in practice this taxonomy is less well defined.

Our objective was to find a language that is easy to understand and is able to express a solution in a very direct and concise manner. To avoid any representation mismatch the data model must be consistent between the application user interface and the programming language. If the prescribed view of geographic information is feature-based, then the programming language must support data access and manipulation using feature structures. Likewise to avoid any problem specification mismatch the style of expression must be consistent between the application user interface and the programming language. If the prescribed view of geo-processing is set-theoretic operations then the programming language must support set queries and set operations on map features.

We believe the geo-relational model is easily comprehended. Users assume that programming a GIS requires manipulating attribute[illegible] or defined map feature sets. In an object-oriented programming environment this type of data and procedural abstraction is easily supported for queries on a single data set. But for compound queries involving several data sets one quickly finds that intricate control constructs and intermediate state information (record numbers, lists of attribute names and values, iterating variables, etc.) are needed. Most users are not accustomed to this programming style, and find it difficult to reconcile the program code with the problem at hand. We believe that a combination of some aspects of functional, object-oriented and rule-based paradigms offers a better solution. An object-oriented interpretation of spatial features provides a simple semantic interpretation of geographical data, all features belong to a themes in a map with appropriate op-

Programming Paradigms	*Procedural Abstraction*	*Data Abstraction*	*Example of Use in GIS*
Logic	-	Meta-rules	*ad hc*
Functional	Functional mappings	Functional mappings	Map Algebra
Rule-Based	Conditional pattern and actions	-	*ad hoc*
Object-Oriented	Sequential, repetition, and conditional branching	Objects, class, inheritance	OpenGIS

Figure 2. Programming paradigms and the types of abstractions employed.

erations on member features. The rule-based paradigm provides a simple mechanism to control program flow. Users can easily grasp the IF-THEN rules and see how it is applied generally. Programmers do not need to construct or navigate between objects, this is inferred from the pattern and action syntax of the IF-THEN rules.

One disadvantage of rules is that they become unyielding and their specification is difficult to understand for nontrivial problems. To simplify the way rules are structured we have explored decision tables. The next section shows how structured rule-sets are organised into a tabular form.

4. Decision Tables

Rule-sets are difficult to interpret for any reasonably sized knowledge base. An alternative technique for representing decision rules is as decision trees (Giarratano and Riley, 1994) or decision tables (Reilly *et al.*, 1987).

The different forms for representing rules can be shown by example. The example describes rules for choosing the best wine to have with a meal.

Given the following rule-set:

IF (main_course is beef) THEN (wine is red)
IF (main_course is fish) THEN (wine is white)
IF (main_course is poultry) AND (meat is light) THEN (wine is white)

C1	main_course	beef	fish	poultry	
C2	meat	-	-	dark	light
A	wine	red	white	red	white

Figure 4: Decision Table

IF (main_course is poultry) AND (meat is dark) THEN (wine is red)

This can be represented in a graph form as a decision tree shown in *Figure 3*.

This can also be represented in tabular form as a decision table shown in Figure 4.

Some of the advantages of decision tables include compactness, self-documentation, modifiability and completeness checking (Reilly *et al.*, 1987). Given that information is stored and viewed in a tabular form in geo-relational databases, it seems fortuitous to represent the rules in a similar form. This presents the user with a very consistent representation of data and procedures.

4.1 Prototype Implementation

The concept of using decision tables for spatial query and modelling was explored by implementing a prototype tool. The tool needed to either interoperate or to be programmed with a GIS that offered object-oriented language features. ArcView (ESRI, 1994) was used because it provided a comprehensive application development environment that included an object-oriented programming language.

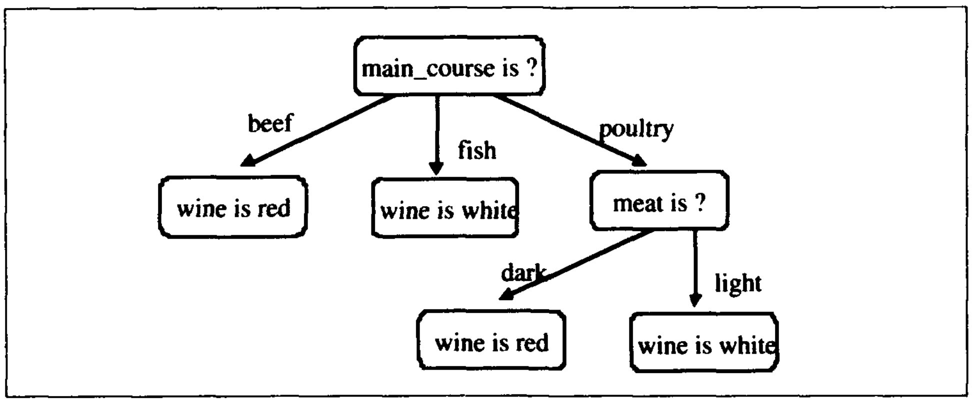


Figure 3. Decision Tree

The organising principle in ArcView is that thematic spatial information is defined and rendered within a geographical portal called a view. A view contains a set of themes all registered to the same geographical space. Each theme represents a defined set of geographic features with their own distinct display characteristics. Each theme corresponds to a geo-relational model of a data source. The geo-relational model (Morehouse, 1985) is based upon a relational data model where recognised tables have one column with values for a spatial domain. A set of these tables each modelling some thematic set of spatial features which share a common geographical extent are the basis of the layered database concept.

The prototype tool was implemented in ArcView using the native programming environment. The algorithm to implement the rule-based approach is a backward chaining inference engine as described in Giarratano and Riley (1994, p:566). Goal objects were matched against the set of features in a theme. The pattern matching was performed on feature-attributes for specified subject clauses and associated values in the columns of the decision table. The syntax adopted was that a table was identified by its thematic name, this was placed in brackets to indicate it represents a free variable that ranges over the set of features in a theme.

For example:

[tree].growth_rate — is a free variable that ranges over the set of features in the "tree"theme with the named at tribute "growth_rate".

Rule inference worked by attempting to match against feature-attributes. The combination of a theme name and attribute domain name identifies feature-attributes in data tables. If a matching feature-attribute could not be found then this was treated as a new attribute derived as part of the inference process. This allowed new derived facts linked to features to be inferred in a natural way. It also relieved the programmer from the burden of setting up variables to hold this state information which was only used during the inference process.

4.2 Example

Two examples of decision tables are described. The first example shows a simple query that may be expressed using an advanced query tool provided within desktop GIS. The second example demonstrates a more complex query that would not be readily represented by any table query tool.

The example is based on a public works problem. Trees located near powerlines need to be periodically trimmed to avoid interference with electrical cables. An application view would include a feature table for tree locations and powerlines.

In the first example a decision table, see Figure 5, is used to express the following query:

check trees within 10 meters of a powerline and have not been trimmed for 2 years.

In the second example a decision table, see Figure 6, is used to develop a more realistic query to account for different growth rates in trees:

check trees within 10 meters of a powerline where the growth from last trim height is now within one meter of powerline height.

Tree height obviously varies over time as a function of recorded height plus growth that has occurred since it

C1	[tree].shape.*DistanceTo*([powerline].shape)	< 10		-
C2	[tree].*SinceTrim*	< 2	-	-
A	[tree].check	true	false	false

Figure 5: Decision Table for first query example.

C1	[tree].shape.*DistanceTo*([powerline].shape)	< 10		-
C2	[powerline].height - [tree].height	< 1	-	-
A	[tree].check	true	false	false

Figure 6: Decision Table for second query example.

C	[tree].type	pine	oak	ash
A1	[tree].growth	0.9 * [tree].*SinceTrim*	0.2 * [tree].*SinceTrim*	0.7 * [tree].*SinceTrim*
A2	[tree].height	[tree].trimHeight + [tree].growth		

Figure 7 Decision Table to deduce tree height.

was last trimmed. The second condition in the decision table specifies a [tree].height which is not a persistent attribute of tress. With pattern matching a decision table is found that lists this attribute (goal) in its action clause. Therefore it is able to infer this information from the decision table shown in Figure 7 and calculate the growth based upon the type of tree.

Growth is given by growth period and a multiplier that varies depending upon the tree type;

growth = $\text{rate}_{\text{tree type}}$ * $\text{peroid}_{\text{since last trim}}$

Note that both feature attributes for [tree].growth and [tree].height are derived for the purpose of the pattern inference and therefore are virtual attributes defined programmatically.

5. Conclusion

The paper has reviewed the different programming paradigms used in computer languages. The goal is to assess what programming paradigm would best suit integration into the user environment of a desktop GIS. We conclude that a combination of a rule-based and object-oriented programming paradigms delivers an easy way for users to perform relatively complex queries and formulate models in a GIS. Researchers have demonstrated that these methods follow the way humans model information and follow human problem solving (Newell and Simon, 1972). We believe that decision tables significantly simplify the layout specifications of rules, and are consistent with the way information is already presented in tabular form in most GIS's.

To support this conclusion we have implemented a prototype query interface to operate with a commercial desktop GIS. Presently the prototype is able to demonstrate the concepts, but has not been developed to the extent that a graphical interface is provided for users to express and execute queries. The next milestone will be to test the tool on a wide range of queries and distribute a robust version of the tool.

6. Bibliography

Abdelmoty A.I., Williams M.H. and Paton N.W. (1993) Deduction and Deductive Databases for Geographic Data Handling. *3rd International Symposium on Large Spatial Databases*, SSD'93, Singapore, pp.443-464

Arentze T.A., Borgers A. and Timmermans H. (1995) The Integration of Expert Knowledge in Decision Support Systems for Facility Location Planning. *Computers, Environment and Urban Systems* 19(4), pp.227-247

Bratko Ivan (1990) *Prolog Programming for Artificial Intelligence*. Addison Wesley.

Davis J.R. and McDonald G. (1993) Applying a Rule-Based Decision Support System to Local Government Planning. In: *Expert Systems in Environmental Planning*, Editors J.R. Wright, et. al. Springer-Verlag, pp.23-45

ESRI (1994) *Avenue - Customization and Application Development for ArcView*. Environmental Systems Research Institute Inc, Redlands, CA.

Egenhofer M. and Frank A. (1990) LOBSTER: Combining AI and Database Techniques in GIS. *Photogrametrmetric Engineering & Remote Sensing* 56(1), pp.919-926

Frank A.U. and Kuhn W. (1995) Specifying Open GIS with Functional Languages. *Advances in Spatial Information Systems*, Proceedings SSD'95, Portland, pp.184-195

Giarratano J. and Riley G. (1994) *Expert Systems - Principles and Programming*. PWS Publ. Co., Boston.

Lowes, D. and Bellamy J.A. (1994) Object Orientation in a

Spatial Decision Support System for Grazing Land Management. *AI Applications* 8(3), pp.55-66

Morehouse S.D. (1985) ARC/INFO - A Geo-Relational Model for Spatial Information. *Proceeding* [illegible]*-Carto* 7, Washington, pp.388-397

Morehouse S.D. (1990) The Role Of Semantics In Geographic Data Modelling. *Proceedings 4th International Symposium on Spatial Data Handling*, Zurich, pp.689-698

Newell A. and Simon H. (1972) *Human Problem Solving*, Prentice-Hall.

Paulson L.C (1996) *ML For Working Programmers*. Cambridge University Press.

OGC (1997) *The Open GIS Consortium*, http://www.opengis.org

Reilly K.D., Salah A., and Yang C. (1987) A Logic Perspective on Decision Table Theory and Practice. *Data and Knowledge Engineering* (2), pp.191-210

Rumbaugh J., Blaha M., Premerlani W., Eddy F., and Lorensen W. (1991) *Object-Oriented Modeling and Design*, Prentice-Hall.

Scarponini P., Clair D., and Zobrist G. (1995) An Inferencing Language for Automated Spatial Reasoning About Graphic Entities. *Advances in Spatial Information Systems*, Proceedings SSD'95, Portland, pp.259-278

Tomlin C.D. (1991) Cartographic Modelling. In: *Geographical Information Systems* - Vol. 1, Editors D. MacGuire, M. Goodchild, and D. Rhind, Longman Scientific & Technical, pp.361-374

Small-scale environments in a fractal landscape

Gordon M.M. Morris
Manaaki Whenua - Landcare Research
P.O. Box 11-052
Palmerston North
New Zealand

Presented at the second annual conference of GeoComputation '97 & SIRC '97, University of Otago, New Zealand, 26-29 August 1997

Many environmental managers face semi-natural landscapes that are complex at various scales. Such is the case of investigating the spread of exotic woody plants in the tussock-grasslands of the Flagstaff-Swampy Ridge near Dunedin, New Zealand. Image processing, supervised classification and analysis was performed on the IDRISI GIS. Three aerial photographs taken in 1975, 1985 and 1990 were digitally scanned, rubbersheeted as ortho-photographs, and used as unintelligent bands. Supervised classification by reclassification was based on the contrast of vegetation patches. Four classes emerged; woody plants, tussocks, grasses and bare ground. Many errors occurred, especially pixel aliasing. Temporal change in the landscape pattern was seen by using fractals, calculated for each patch type in a raster environment. The increasing fractal dimension for the tussock patches over time corresponds to an invasion of woody plants and subsequent fragmentation. Ground truthing supported this finding, showing that the heterogenous landscape is linked by small area environments that offer safe-sites to plants, especially pig rootings. Future IGIS will combine knowledge about a plant's characteristics linked the landscape. This will allow a greater understanding of the spread and establishment woody exotic plants in New Zealand's landscapes.

Key words; environmental heterogeneity, exotic woody plants, fractal dimension, GIS, IDRISI, pig rooting, raster, safe-sites.

1. Complex environments and patch dynamics.

Seen by some as a threat to the natural flora and fauna, to others a 'natural' part in the evolution of the landscape, invading plants in New Zealand have been a continuing topic of debate.

Interactions between environmental variables are often complex, especially over semi-natural tussock landscapes. Traditional ecological methods and transects are often too simple to explain patterns, especially where wilding conifers are found in New Zealand's high country. Managers of these lands face a wealth of often conflicting knowledge. The spread of wilding exotic conifers can be identified as originating from "take-off sites" (Ledgard and Crozier, 1991). These areas are favourable for growth, and subsequent dispersal of wind blown plants. Analysis of natural patterns needs an unbiased approach to pattern analysis from accessible imagery, especially where woody species spread along natural boundaries.

2. Patterns and analysis.

Suites of aerial photography allow cost-effective images for showing detailed change in the environment. They visualise theories of landscape ecology that see the environment as a mosaic-like landscape pattern of individual, but interconnected, patches. These patches are communities or species assemblages are surrounded by a matrix of vegetation of distinctive structure or composition (Forman and Godron, 1981). Patterns in a landscape are representative of a non-uniform resource distribution in the

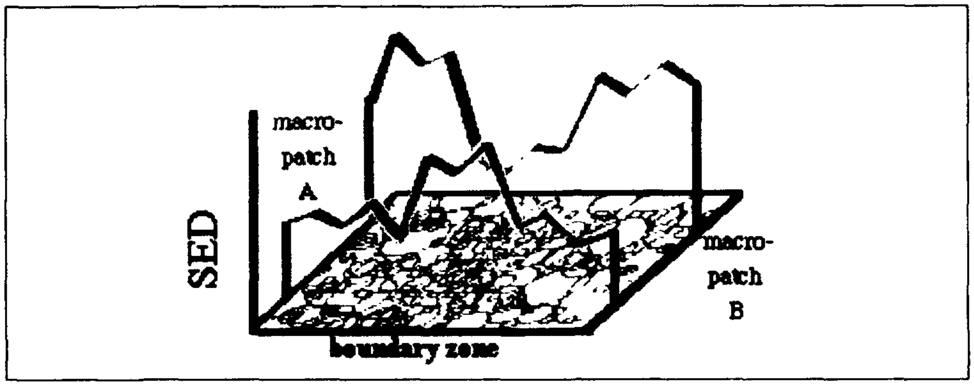


FIGURE 1. Hypothesised spatial arrangement between macro-patches.

landscape. The patch boundary can be defined in terms of a gradient between two neighbouring macro-patches, where the boundary is the locus of points that exceed a specified threshold (Musick and Grover, 1991). Ludwig and Cornelius (1987) documented change along an environmental gradient, using a split-window gradient analysis. They found "discontinuities;" areas of extreme differences in the hypothetical positions of patches along the gradient. Orlóci and Orlóci (1990) identify that raw populations are being compared, however their differences from the expected Squared Euclidean Distance values would represent mosaic-like arrangements between macro-patches (Figure 1). In vegetation studies, these may correspond to contrasting vegetation zones.

Although the split-plot may not coincide with the actual environmental or vegetational edge, Hardt and Forman (1989) show that the shape of an edge is critical for the recruitment of woody seedlings. This reflects varying scales of activity, both abiotic and biotic. They stated that a boundary has a thickness even when the edge is discrete. Hence 2-dimensional studies are required to identify boundary zone dynamics. Geographic Information Systems (GIS) offers the opportunity to use images for analysis, representing the environment at fine scales; thus allowing a holistic perspective.

The holistic environment as hypothesised in Figure 1 comprises smaller entities; and is only homogenous at the smallest scale, excepting fuzzy boundaries. This is shown by Schaller (1994) by the application of landscape units to a GIS, that an effective overlaying of patches can best be achieved by determining the 'smallest common unit'. Without reference to heterogeneity, this shows that the smallest part of the environment has its own unique qualities of measured heterogeneity and small-scale processes. This is especially applicable in a GIS approach to analogies of a complex environment. This does away with the need to define a boundary in terms of ecotone or ecocline. Instead, it refers to the characteristics of the patches themselves as part of a combined landscape. Greater emphasis is given to the landscape and individual plants. By measuring patch perimeter to patch area, fractals offer an independent guide for landscape ecology. From work in soil mapping, Burrough (1983) shows low fractals as "short-range variations" and high fractals as long-range hieratical effects. Forman (1997) also described hypothesized values of patch shape reflecting patch stability. Simple shapes with low fractal values were stable, higher fractals represented a complex environment. These can be used to measure the opportunities for woody plant expansion and their effects along an ecological boundary, by using a detailed remote sensing study in a GIS environment.

3. Methods.

The study area was in the headwaters of Nichol's Creek, Flagstaff-Swampy Ridge, near Dunedin, New Zealand (Figure 2). Vegetation is predominantly *Chionochloa rigida* (snow tussock), with *Phormium cookinum* (flax). Patches occur of

FIGURE 2 The study area, looking south (Study site outlined in white).

invading woody plants; natives *Cassinia vouvilliersii* and *Leptospermum scoparium*, and exotic *Cytisus scoparius* and *Ulex europaeus*. The area, now part of a water reserve, was once grazed and has been oversown with many exotic grasses. IDRISI, a raster-based GIS, was used for image enhancement and classification of vegetation patches, and calculation of the fractal dimension in a raster environment.

Enlargements of contact prints from aerial surveys in 1985 and 1990; and a bromide copy from 1979 were scanned in at 360 dots per inch on a flat-bed scanner, then imported into the IDRISI environment as 256 grey-scale TIFF files. Mean and median filters were used to remove random noise in the images from silver nitrate crystals in the enlarged photograph that showed as high and low values in homogenous areas. Images were rubbersheeted to produce aerial ortho-photographs, which gave independent bands. Images were matched to the 1990 image due to the number of points that could be identified with reasonable certainty. Prominent points such as fenceposts, crowns of individual trees, patches of bare ground were used as control points between images. A bilinear quadratic movement was used to speed processing time. Points were clustered along the boundary to give a higher accuracy where the most change was expected occur. The reformatting also gave each set the same number of columns and rows. The area used for the analysis was an approximate 500 meter square subsampled area, taken to minimise the large distortion between bands. A larger contrast in grey-scale values was made by stretching the band to give 256 values. Reclassification gave four vegetation classes: woody plants, 'tussock', 'grass', and 'bare' ground. The decision points for this supervised classification are given in Table 1. Patches were identified with the GROUP function, matching like pixels in the same unique identifier, including those on the diagonal.

To overcome the effects of geometry of pixels either singly or as groups, Olsen *et al.* (1996) modify the fractal dimension from Peitgen and Saupe (1988) by avoiding the regression of the patch perimeter (P) to the patch area (A). Olsen *et al.* (1996) then calculated the constant of proportionality from Equation 1, where a single pixel (with four sides) is the simplest case. Equation 1 solved for k, a single pixel with area 1 and perimeter 4 as the simplest case. The constant of proportionality for the cell is obtained as k = 4. The fractal dimension used for this research for calculating the fractal dimension of a patch in a raster had the constant subsitutated in and the log taken (Equation 1).

$$D = 2 * \ln (P/4) / \ln (A) \qquad \text{Equation 1.}$$

TABLE 1. Decision points used in the supervised classification of each image

	classification types			
image	**woody**	**tussock**	**grass**	**bare**
1979	1 - 114	115 - 212	213 - 255	not classified
1985	1 - 122	141 - 186; 208 - 219	122 - 140; 186 - 207	220 - 255
1990	1 - 151	152 - 176	177 - 199	200 - 255

TABLE 3. Mode of the fractal dimension (D value) in the study and sub-sample sites

	Study site patch type				Sub-sample patch type			
image	woody	tussock	grass	bare	woody	tussock	grass	bare
1975	1.47	1.73	1.51	———	1.29	1.61	1.54	———
1985	1.41	1.56	1.62	1.40	1.37	1.62	1.78	1.34
1990	1.35	1.70	1.58	1.38	1.38	1.38	1.55	1.46

Single pixel polygons were removed, and like pixels grouped together as patches with unique identifiers. The area and perimeter of each patch was calculated, and the fractal dimension (the D value) was found as for Equation 1 by overlaying images.

A subset of 1000 by 1000 pixels randomly placed within the study site was taken to identify local effects of scale on the fractal dimension; single pixels were not removed. The results from this are discussed after analysis of the study site.

4. Analysis.

A reclassification of grey-scale values was used to distinguish between patches. This gave a better separation between classes in the bands than signatures did, due to the overlapping range of values within each signature file. Where values overlapped there was a high chance that a pixel was assigned to the wrong class. This can be seen as 'noise,' for example tree crowns can be identified by their lightness, and are classed as 'tussock.' There may also be misclassification, where flax plants are seen as woody species because of their low grey-scale values.

Some categories had to be added together: for example, pine trees and native bush as 'woody', and 'grassed' areas with bare ground as 'grass' in the 1979 image. Stretching the 1979 band to the full 256 grey scale values did not increase the contrast due to areas of high reflectance; it is likely that both 'grass' and 'tussock' values have bare ground included in them. This is not a problem except where species specific information is required, and where textural values change with position and aspect.

The modal frequency of pixels indicated the fractal dimension of the macro-patch. Tussock and 'grass' patches enclosed by invading woody plants show a decrease in their fractal dimension. Such patches are small, and often regular. This may indicate that they are individual tussocks, stable areas of tussock, or patches unsuitable for woody plants. By 1990 the tussock macro-patch has broken up into smaller patches with a similar, and lower, fractal dimension. As the environment fragments, the D value increases. When the patch is reduced in size, patches with a lower value remain. This shows that fractals indicate the fragmentation of the landscape.

Smaller patches of tussock show lower fractal dimensions (approx 1.40) and occur around the outside of the macro-patch towards the woody macro-patch. This may indicate that they are constant within the landscape - as they are in the area of a patch of *Chionochloa conspicua*. However, the tussock macro-patch also crosses the main track in part and must be used with caution when seen as a whole as there will be pixel aliasing in these track areas. The decrease in the D value for woody plant patches occurs as the area becomes greater, with the macro-patch joining 'woody' patches in the tussock macro-patch. The value of D is significantly different from the tussock macro-patch, and separates the two macro-patches in terms of spread.

The patches within the woody macro-patch are noise from errors in classification, but their effect is minimal. However, some heterogeneous patches are classed as both 'tussock' and 'grassed'. This suggests the underlying heterogeneity within both of these classes. The fractal dimension of grassy patches in the woody macro-patch is 1.30; a low D value perhaps indicating a stable influence within the environment. The presence of this patch can be seen in aerial photographs since 1949. However the 'grass' patches, which surround many tussock patches, have fractals averaging 1.39, but reflect fragmentation, especially in the 1990

image. They are the primarily places where woody plants invade. This is similar to the 1990 'bare' patches in an area of pig rooting, which has a D value of 1.30.

Pig rootings offer safe-sites (Harper, 1977) for plant growth in the environment. In these disturbed areas, seedlings are protected from drying winds, gives insulation from cold, have a constant nutrient supply, and a reduction in competition by other species. Gorse seedlings, from the seed bank, are quick to occupy bare ground and form close knit patches. This is an important interaction at the small-scale and an important mechanism for disturbance in this area.

Other identifiable patches form to make the track, which is classed as 'bare', 'grassed', and 'tussock'. Its fractal ranges from 1.30 to 1.40 in the 1979 and 1990 images, but is 1.60 in the 1985 image. This high value is due to its irregularity with its length, or misclassification from shadows along its edge. Otherwise, its low value of D corresponds to a stable patch within a landscape, even though they are not linked together as the same patch type.

The randomly placed sample of the site 1000 by 1000 pixels assessed the change in scale on the fractal dimension of the macro-patches. Table 3 shows a reversing trend of modal values for the fractal dimension to the study site. This occurs from the local effects (Feder, 1988) of re-sizing the woody and tussock macro-patches; and the straight lines of the edge of the image, which also give simpler shapes.

Some values of D for the study site were not included in the analysis, tending to be complex patches greater than 1.61 (Figure 3). This is the effect of the mean and median filters smoothing complex shapes during image enhancement. However, without filtering high impulse noise would increase; as will the complexity of the method.

5. *Data quality.*

Goodchild (1994) argues that although remote sensing and GIS are applicable to vegetation analysis they vary as to their integration. Cartographic boundaries are often seen as discrete lines or homogenous areas of constant width between distinctive spatial units. Small areas may also be smaller than the minimum mapping distance and, as with most cartographic drawing, are subject to generalisation. The method allowed a landscape boundary between two ecosystems to be a heterogenous area composed of patches, smaller than the minimum mapping distance for conventional map adding to the detail of the study.

However, the scale of investigation in a raster is determined by the pixel or grid size, so the smallest feasible scale of investigation remains uncertain. There must be a balance between too much information, and over generalisation. The orientation of the raster is also of prime importance, especially at larger scales where the grid must be orientated with the boundary of the study area in mind. This is because a boundary and a sampling grid must be similarly orientated (Burrough, 1987). In that important respect, the topology of the raster system may not match coordinates which are plotted close together. Thus the fractal dimension calculated for square metres is different from cell values. Further, objects may not fit into the grid of a square raster especially where the object is too small

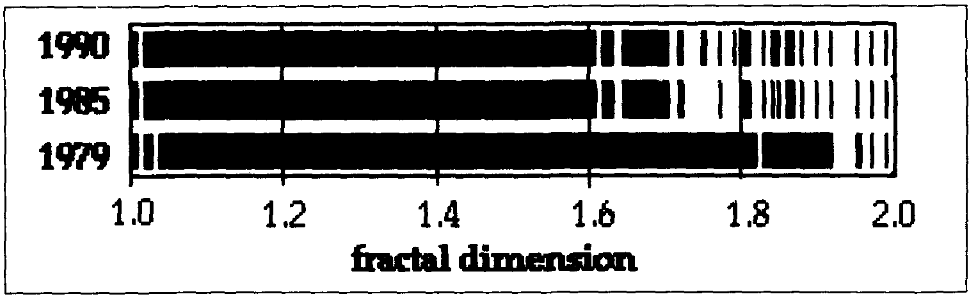


FIGURE 3. Presence/absence (black/white) of fractals, for all patch types.

for the pixel, so the pixel becomes an alias of the true feature (Gahegan, 1994). Objects that are able to be captured may not be positively identified. This is the case of pixel aliasing, where pixels do not represent the object they describe. This occurs where large-area pixels do not include the classes of information, or the decision points in the reclassification are incorrect.

5.1 Errors in classification.

Ground truthing shows that some areas may appear as woody plants, but are flax or the small shadows of large tussocks; and that 'grassed' areas (as seen by their lighter tones) may include some tussocks, especially on sunlit slopes. This means that the error between images is not constant. Differences in shadow angle, light quality between images, and the reflectance of the ground and vegetation meant that classifications were not interchangeable between bands, and errors would occur over the study site.

A large source error in data capture comes from the use of only one bandwidth per data set (visible light as seen by 256 tones of black and white). With the reclassification remote sensing, some pixels will be reclassed incorrectly. This can be seen by the number of 'tussock' patches with in the woody macro-patch. The photographic paper may not be as responsive as a digital signal, giving a systematic type-sample. Where the photo is lighter, as on the northern side of a hill, it also has higher reflectance values than southern sides. These 'hotspot' points occur where the photograph's angle varies in relation to the sun's angle, a problem with ortho-photographic images that have not been corrected to the nadir (ie. solar angle directly over head). The reflectance of pine trees may be up to five times the original reflectance value (Dymond, 1996), and would be even greater for bare land. Future Integrated GIS (I...) will incorporate Triangulated Irregular Networks (TIN) that combined with co-ordinates from sub-metre Global Positioning Systems (GPS), tilt and tip displacement from the camera would be removed and allow for differences in the topography (relief displacement) to be removed. Further, for analysis by the supervised classification, a weighted index of classified grey-scale values for the various slope aspect and angles would reduce errors from shading effects and bright illumination when using reclassification.

This also shows in the reclassification of grey-scale values for a single tree (Figure 4). The vector line shows the approximate position of the plant, the dark area below this is the plant's shadow. A histogram of the spectral response for this plant shows the range of pixels with grey-scale values [indicated along the x-axis] from 58 to 188, with the higher numbers being lighter. Much of the classification from analysing the histograms of training sites would overlap with other sites.

Although some generalisation may be made by this method, the error is still less than that by individually assigning each small-area pixel to a class based on a training site. The minimum mapping distance would also require eight pixels to surround a 'patch' of one pixel. Ground truthing validated most of the supervised classification results and emphasised the importance of small-scale.

The test site for the fractal dimension shows that the scale of resolution is important; the analysis of macro-patches

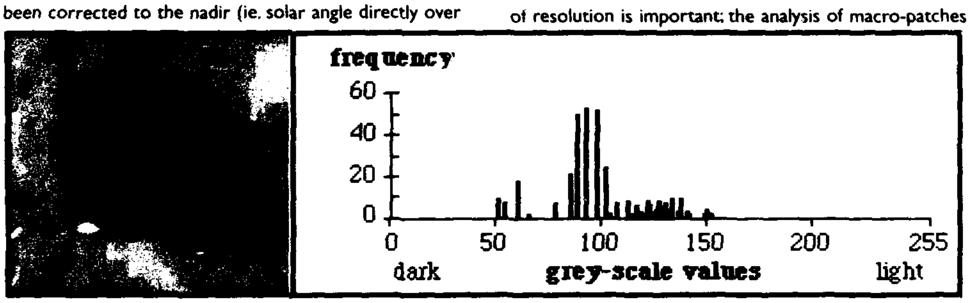


FIGURE 4. Histogram of a single tree from the 1990 band (outlined left)

should identify largest patch and use that as the boundary for further study. Thus landscape studies with fractals from raster images should be from the largest area and resolution possible. This will also avoid errors that occur from the straight edge of the sub-sampled area being included in a patch's perimeter.

6. Fractals and vegetation fragmentation.

Figure 5 shows the tussock macro-patch from 1979 to 1990. Although entirely connected, this patch has been pixel thinned (1 in 10) for display. The woody plants invade into the 'grassed' and 'tussock' areas, opening up the tussock macro patch. They overtop the tussocks when they expand. This increases the 'woody' area and decreases the perimeter to area ratio. Expansion by woody plant growth has meant that the boundary between the two macro-patches has become in-filled. This smooths out the complex area and decreases the value of D for woody plants.

Miller (1994) noted that heterogeneity in the environment could be seen from the apparent "holes" in a patch. This heterogeneity and fragmentation is shown by patch isolation, where the macro-patch becomes opened up from within (Merrin and Wegner, 1992); smaller units become separated from one another as connections within the macro-patch disappear. These patches have lower D values, and may be stable as individuals and patches resistant to invasion, unlike the hypothesis of Forman (1997). Thus what the contrasting fractals show is the dynamic fragmentation of the tussocks from the invasion of woody plants, on the Flagstaff-Swampy Ridge.

7. Conclusions.

For research into the landscape, ortho-photographs and fractal analysis are inexpensive and highly effective tools for the landscape planner to build up a database of the area using GIS. As patches in the present study were defined by the canopy cover, the fractal dimension does not show how the patches are occupied. The fractal dimension describes landscape boundary dynamics, where interactions between macro-patches are of interest. If small-scale environmental conditions for survival of an invasive plant are available from the interactions within the environment, it will tend to be in a landscape boundary zone. Small-area analysis by a GIS with a plant database would identify the mechanisms of those forces that promote change in the landscape, as would an integrated DTM-wind field to show take-off sites. An overlay of the fractal dimension for functional and measurable physical properties may be a profitable area for future research.

Acknowledgment.

This paper is from a thesis to be submitted in partial requirement for the Master of Science Degree, University of Otago, Dunedin, New Zealand. I wish to sincerely thank my supervisor, Professor Peter Holland for his encouragement, and the two anonymous referees for their constructive comments. *Mo Rich Mo Dhuthaigh.*

FIGURE 5. Fragmentation of the tussock macro-patch (shaded).

References.

Burrough, P.A. (1983) Multiscale sources of spatial variation in soil. I. The application of fractal concepts to nested levels of soil variation. *Journal of Soil Science* 34:577-597.

Burrough P.A. (1987) Spatial aspects of ecological data. In Jongman, R.H.G., ter Breek, C.J.F. and van Tongeren, O.F.R. eds. *Data analysis in community and landscape ecology.* Pudoc, Wageningen, the Netherlands 1987, pp213-251.

Dymond, J. (1996) Vegetation viewed from above. *Information In Formation*, a Landcare Research New Zealand Ltd. periodical, pp2-3.

Eastman, J.R. (1995) *IDRISI for Windows: users guide version 1.0.* Clark Labs for Cartographic Technology and Geographic Analysis. IDRISI Production, Clark University.

Feder, J. (1988) *Fractals.* Plenum Press, New York.

Forman, R.T.T. (1997) Land mosaics: the ecology of landscapes and regions. Cambridge University Press.

Forman, R.T.T. and Godron, M. (1981) Patches and structural components for a landscape ecology. *Bioscience* 31:733-740.

Goodchild, M.F. (1994) Integrating GIS and remote sensing for vegetation analysis and modelling: methodological issues. *Journal of Vegetation Science* 5:615-626.

Hardt, R.A. and Forman, R.T.T. (1989) Boundary form effects on woody colonisation of reclaimed surface mines. *Ecology* 70:1252-1260.

Harper, J. (1977) *Population Biology.* Academic Press, New York.

Ledgard, N.J. and Crozier, E.R. (1991) *Guidelines for the control and management of wilding pines in the Canterbury high country.* Forest Research Institute Contract Report, FWE 914.

Ludwig, J. A. and Cornelius, J.M. (1987) Locating discontinuities along ecological gradients. *Ecology* 68:448-450.

Merrin, G. and Wegner, J. (1992) Local extinctions, habitat fragmentation, and ecotones. In Hansen, A.J. and di Castri, F. eds., *Landscape boundaries and consequences for biotic diversity.* Springer-Verlag, New York, 1992, pp150-169.

Miller, D. (1994) Coupling of process-based vegetation models to GIS and knowledge-based systems with reference to vegetation change. In Worboys, M.F. ed. *Innovations in GIS 1.* Talyor and Frances, London 1991, pp241-250.

Musick, H.B. and Grover, H.H. (1991) Image textural measures as indices of landscape pattern. In Turner, M.G. and Gardner, R.H. eds., *Quantitative methods in landscape ecology: the analysis and interpretation of landscape heterogeneity.* Springer-Verlag, New York, 1991, pp77-103

Olsen, E., Ramsey, R.D. and Winn, D.S. (1996) *A modified fractal dimension as a measure of landscape diversity.* Unpublished manuscript.

Orlóci, L. and Orlóci, M. (1990) Edge detection in vegetation: Jornada revisited. *Journal of Vegetation Science* 1:311-324.

Peitgen, H.O. and Saupe, D. (1988) *The science of fractal images.* Springer-Verlag.

Richards, J.A. (1993) *Remote sensing digital image analysis: an introduction.* Springer-Verlag, Berlin.

Schaller, J. (1994) Geographic information systems and ecosystem models as tools for watershed management and ecological balancing in high mountain areas: the example of ecosystem research in the Berchtesgaden area, Germany. In Price, M.F. and Heywood, D.I. eds., *Mountain environments and geographic information systems.* Taylor and Francis Limited, London, 1994, pp43-58.

Posters

First Experiments in Neural Network Mapping

Robert J. Abrahart
School of Geography, University of Leeds,
Leeds LS2 9JT
United Kingdom
Tel: +44 - (0)113 - 2336750
bob@geog.leeds.ac.uk

Presented at the second annual conference of GeoComputation '97 & SIRC '97, University of Otago, New Zealand, 26-29 August 1997

1. Introduction

Neural networks are now accepted tools in many areas of business and science; comprising an important group of powerful emergent data-driven technologies, sometimes described as a "solution looking for a problem". Reasonable simulation programs are now available in the form of commercial and public domain packages - for both UNIX and PC platforms. These modern computer-based solutions have various distinct real-world operational uses; their practical implementation can be achieved in several alternative modes (pertaining to their association with existing ideas and methods); and they possess a number of significant computational advantages that can be profited from e.g. "intelligent" data analysis and/or modelling, high-speed information processing, and robust data handling/error tolerance. However, within the geosciences, the application of neural networks has thus far tended to focus on the mundane replication of existing equation-based tools - or on solving problems that are of a simplistic or otherwise straightforward nature e.g. satellite image classification. At the heart of this bottleneck lies a fundamental belief in existing solutions and an unwillingness to explore beyond that which is known and trusted.

Much of our existing available geographical data resides as grid-based maps or models within a raster GIS; and satellite information continues to be supplied in a similar format - albeit spread across several different spectral bands - at an ever increasing rate. In order to cope with the anticipated growth in demand for future geographical products and solutions - that will be required from such data - there is a pressing need to maximise our effort towards devising new and/or alternative approaches to the problematic task of storing, manipulating, and processing spatial information. Neurocomputing offers one possible answer, and to help foster "increased awareness" of potential neural network solutions within the geographical sciences, three simple experiments have been carried out in an initial attempt to explore the opportunities associated with employing neural networks for replicating, improving, and creating raster-based products. In each case the proposed solution demonstrates the capabilities of this novel approach to tackling otherwise complex mapping and map-based-modelling problems. The results of this exercise are best visualised in graphical form using appropriate maps and diagrams - which are provided here and on the accompanying poster.

2. Replicating Multiple Maps

New methods are needed to overcome the two related spatial data handling problems of information storage (enormous volume requirement) and data retrieval (rapid access requirement). If raster maps, either alone or in related groups, could be replaced with neural network models - the storage space requirement would be reduced to minuscule levels and information processing operations could switch from file-based data retrieval (slow) to chip-based data computation (fast) procedures.

Brakensiek & Rawls (1983) in their work on the use of

infiltration procedures for estimating runoff produced a number of "soil texture look-up charts" - from which can be obtained various parameters associated with the "Brooks-Corey Soil Water Retention Equation" (Brooks & Corey, 1964) and "Green-Ampt Infiltration Equation" (Green & Ampt, 1911). These charts were developed from simulations based on c. 5,000 soil data records and are said to represent average soil conditions prior to a particular agronomic practice. Each chart comprises a three dimensional surface, plotted as a limited number of isolines in soil texture space, and has a triangular format: x-axis being percentage clay, y-axis being percentage sand, and z value being the required soil parameter. Different charts were produced for various different organic matter percentages, with each "soil parameter and organic matter percentage combination" comprising a set of four triangular diagrams, wherein each triangle represents the percentage porosity change associated with a different level of surface compaction. Two soil parameters were selected for modelling viz. effective porosity [cm3 cm-3] and saturated hydraulic conductivity [cm hr-1]. The relevant charts were those pertaining to the 0.5% level of organic matter. There were eight charts in total - comprising four triangular diagrams for each of the two soil parameters - with each triangle representing a different level of porosity change associated with surface compaction (0%, 10%, 20%, and 30%). The eight isolines charts were digitised using ARC/INFO. To increase the number of significant figures and facilitate later integer-based processing the isoline values for effective porosity and saturated hydraulic conductivity were multiplied by 10,000 and 1,000 respectively. The digitised vectors were converted to node-based point data and all points reflected in space using a bespoke awk script - thus extending their actual borders - to help minimise the production of spurious edge effects in subsequent interpolation operations. Eight interpolated raster surfaces were constructed from the expanded point data in GRASS (Geographic Resources Analysis Support System) using "regularised spline fitting with tension and smoothing" (Mitasova & Mitas, 1993a & 1993b). Each final map comprised a raster grid of 2000 x 2000 cells - with the original triangle being located in the upper lefthand corner of the central 1000 x 1000 square block. 5,000 random point samples were taken from within the area of each original triangle on each map, the co-ordinates for this operation being held in a random lookup table, generated from a uniform distribution - with the extracted data being written to file. The final data contained 20,000 patterns, comprising 5,000 points for each level of porosity change, and with each pattern containing five variables: percentage sand; percentage clay; percentage porosity change; effective porosity; and saturated hydraulic conductivity. All five variables were then subjected to linear normalisation between zero (lowest possible value for that variable in the dataset) and one (highest possible value for that variable in the dataset). The normalised file was split into two equal data sets; one for training the network, the other for split-sample validation purposes.

The Stuttgart Neural Network Simulator was used to construct a two-hidden-layer feedforward network with a 3:12:12:2 configuration and with all appropriate connections enforced. The input nodes were for percentage sand, percentage clay, and percentage porosity change. The output nodes were for effective porosity and saturated hydraulic conductivity. Network training was undertaken using "backpropagation without momentum". The learning rate was reduced according to a sliding scale at pre-set intervals and the network was observed to converge in a smooth and uneventful manner. Training was stopped at 30,000 epochs. Error reduction was observed to be almost non-existent after 30,000 epochs indicating broad-scale convergence. The average final sum squared error per output node was just over 0.36 normalised units. At each point of change in the learning rate both training and validation datasets were passed through the trained network in its non-training mode and network output plotted against model output. In all instances the two plots were quite similar - which is indicative of good generalisation and modelling. Scatterplots for the validation data at the final stage of the learning process are reproduced here in Figures 1 and 2. Whilst some almost insignificant discrepancies can be seen to occur in the uppermost and lowermost sections - in each case the end product other-

wise exhibits a near perfect agreement between observed and predicted values. This experiment altogether demonstrates the unharnessed potential of using neural networks to model smooth map surfaces.

3. *Improving Existing Maps*

All maps have errors - amongst the worst offenders being statistical surfaces constructed from sparse and irregular point distributions. Netherthelss, some maps will of course be more accurate than others, and the contention

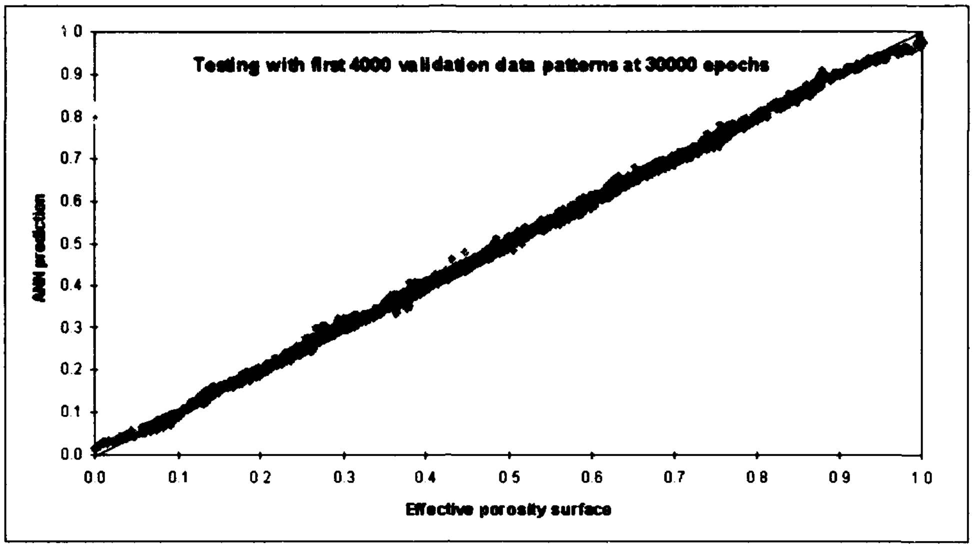


Figure 1: Modelling effective porosity values derived from soil texture look-up charts with an artificial neural network. Scale is in normalised units. Line of perfect agreement drawn in black.

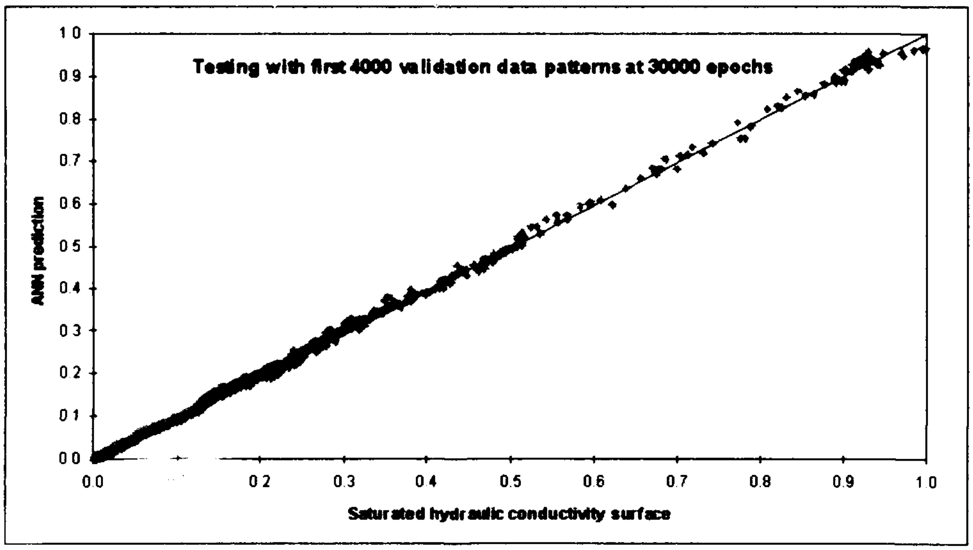


Figure 2. Modelling saturated hydraulic conductivity values derived from soil texture look-up charts with an artificial neural network. Scale is in normalised units. Line of perfect agreement drawn in black.

here is that good continuous distributions can be used to filter out (legitimise) the numerous spatial inconsistencies that exist within their poor or erroneous counterparts. It is quite feasible for a neural network to model all possible domains within a spatial database including those important relationships that exist both within and between the various individual components e.g. if a network is trained with a combination of locational (relative or absolute) and multiple environmental data then the solution surface will perforce be all embracing. The neural network would use otherwise unknown relationships that exist within the spatial data to form its model, thus providing a multi-source holistic tool for predicting the spatial distributions of environmental phenomena, which could operate at the level of an individual cell within each raster grid. This process would provide a robust error-tolerant multi-dimensional non-linear solution to what is otherwise a difficult modelling task; and at the same time create a mechanism that could be used to recognise and remove tangible inconsistencies.

An interpolated map of long term mean annual rainfall (LTMAR) for the period 1961-1990 was constructed for North West England (13,000km^2) based on raingauge data. This is a large region that extends from Buxton in Derbyshire (south-east) to Carlisle in Cumbria (north-west), and comprises a diverse area encompassing the north west seaboard, the lowlands of the Lancashire Plain, and the major upland areas of England (including the Lake District and Pennines). There are 1,384 raingauge sites in this region and most of them are in the lowlands (Figure 3). In the uplands, there are few raingauge sites, and their spatial and elevational distribution is quite uneven. The information collected at these sites is used for constructing interpolated surfaces of average values, and these surface values are in turn used to calculate water balances for reservoired upland catchments, which provide potable water for cities such as Manchester. The LTMAR surface was generated to a 50m grid, using gaussian kriging, in ARC/INFO. It contained a lot of internal smoothing, often across large areas where there had been little or no original input data, and had serious edge problems. Additional surface information was available in the form of an O.S. "Land Form Panorama" 50m x 50m raster digital elevation model, from which slope and aspect values for each cell in the raster grid were computed, using standard GIS tools. A variance surface (confidence measure) was also generated as a natural output of the kriging operation. In this investigation the aim was to model the general relationship be-

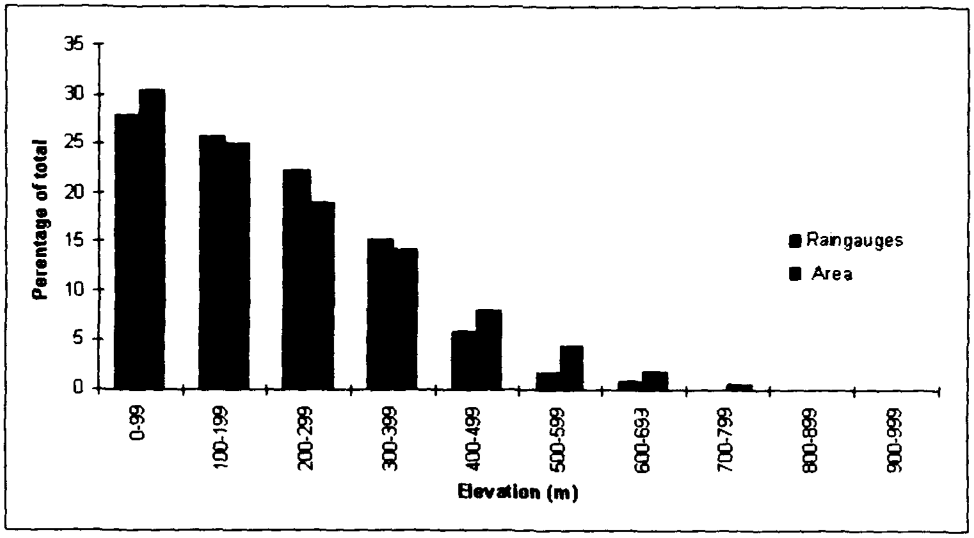


Figure 3: Elevational distribution of raingauge sites in North West England.

tween a combination of topographic and locational factors (independent variables) and the kriged LTMAR map (dependent variable) - based on the recognised influence of terrain on rainfall - with the ultimate goal of creating a mechanism that could filter out (legitimise) the most obvious inconsistencies. 16,000 random points were sampled from the five raster layers in the GIS database (elevation, slope, aspect, variance and LTMAR maps), the co-ordinates for this operation being held in a random lookup table, generated from a uniform distribution - with both sampling coordinates and extracted data being written to file. Given the circular nature of "aspect" these values were transformed into their sine and cosine equivalents. All eight variables were then subjected to linear normalisation between zero (lowest possible value for that variable in the GIS database) and one (highest possible value for that variable in the GIS database). The normalised file was split into two equal data sets; one for training the network, the other for split-sample validation purposes. The Stuttgart Neural Network Simulator was used to construct a two-hidden-layer feedforward network with a 7:18:18:1 configuration and with all appropriate connections enforced. The input nodes were for easting, northing, elevation, slope, cos(aspect), sin(aspect), and kriged variance. The output node was for kriged LTMAR. Network training was undertaken using "backpropagation without momentum". The learning rate was reduced according to a sliding scale at pre-set intervals and the network was observed to converge in a smooth and uneventful manner - albeit with sharp drops at each change in the learning rate. Training was stopped at 30,000 epochs. Error reduction was observed to be almost non-existent at this point indicating broad-scale convergence. The final sum squared error for the output node was just over 2.67 normalised units. At each point of change in the learning rate both training and validation datasets were passed through the trained network in its non-training mode and network output plotted against model output. In all instances the two plots were quite similar - which is indicative of good generalisation and modelling. A scatterplot for the validation data at the final stage of the learning process is reproduced here in Figure 4. Throughout most of the plotting space the output data exhibits a modest spread of values; there are no major outliers, and the general trend has a close association with the line of perfect agreement - with values both above and below it - thus providing further evidence

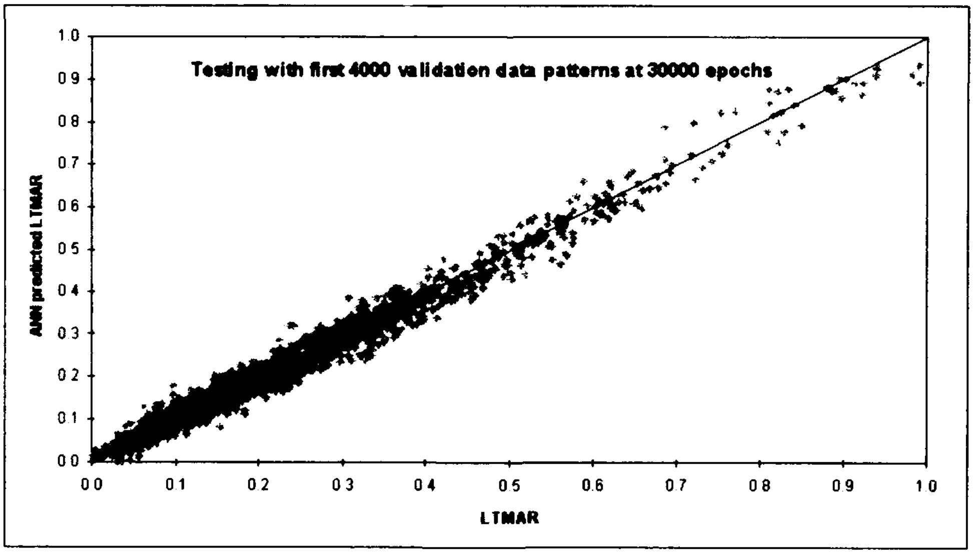
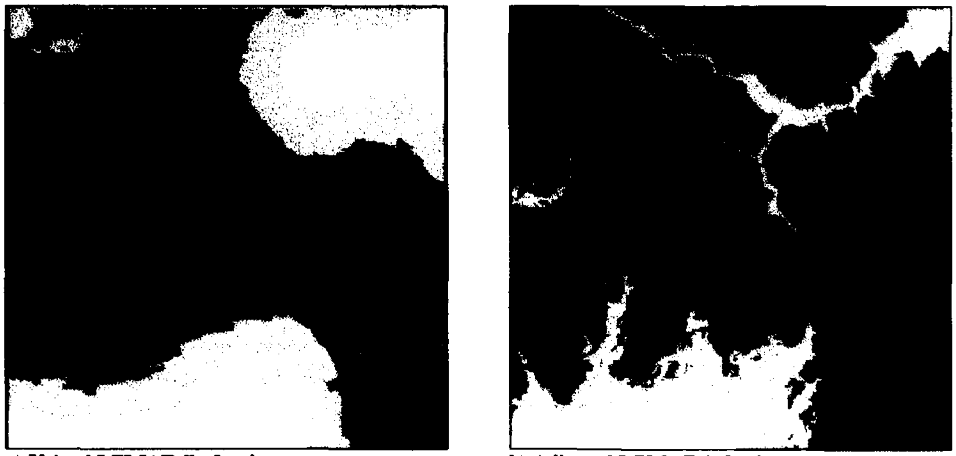


Figure 4 Modelling Long Term Mean Annual Rainfall in North West England with an artificial neural network. Scale is in normalised units. Line of perfect agreement drawn in black.



a) **Kriged LTMAR [before]**
Range = 1150mm (light grey) to 1566mm (black)

b) **Adjusted LTMAR [after]**
Range = 1123mm (light grey) to 1509mm (black)

Figure 5 Long Term Mean Annual Rainfall maps for the Rochdale-Todmorden area

of good modelling. In the uppermost sections of the plot a slight but nethertheless consistent underestimation is observed. These high values are concentrated in one small geographical area - the Lake District - which is therefore perhaps not modelled to the same standard as the rest of this region.

Two representative windows were next selected and extracted from the GIS database, converted into the required format, and passed through the trained network viz.: a mountainous area with high absolute adjustments (Lake District, 267km²) and an inland area with low absolute adjustments (Rochdale-Todmorden, 192km²). The outputs from this exercise were shipped back into the GIS thus creating two corrected LTMAR maps that could be used for visual and statistical analysis. Before and after maps are provided for the Rochdale-Todmorden area in Figure 5. These two maps exhibit a similar range indicative of detailed adjustment and fine-tuning (original, 1150-1566mm; corrected, 1123-1509mm) and the corrected map exhibits numerous minor modifications throughout - the extent of these changes ranging from -252 to +199mm. Moreover, in accordance with theoretical knowledge about the relationship between elevation and rainfall, the final map now better mimics the elevation surface (high positive correlation, + 0.89). There also remains a reasonable positive correlation between the original and corrected LTMAR maps (+0.58) which is a measure of the level of adjustment that has been made. It is logical to assume that a good result would produce a positive "middle of the range" statistic since a high correlation would indicate insufficient alteration (overfitting) and a low correlation would indicate excessive adjustment (underfitting). The results also show a marked decrease in LTMAR for high rainfall values and a marked increase in LTMAR for low rainfall values (high negative correlation between original LTMAR and neural network adjustments, -0.72) which is instructive. This experiment altogether demonstrates the *unharnessed potential of using neural networks to form* complex spatial models at the regional scale - for error trapping, surface adjustment, and data investigation purposes.

4. *Creating New Maps*

Derived products created using either standard or bespoke GIS functions are now commonplace e.g. slope, aspect, and flow accumulation maps generated from digital elevation models. Nethertheless, several standard GIS algorithms are now criticised in the literature as being grid-square

case-intolerant, and commensurate with these simplistic or inappropriate algorithms being applied in an unskilled manner - there is often a deleterious knock-on effect (Zhou et al., 1997). Moreover, standard modelling practice requires one to choose between a limited number of alternative system-dependent equation-based strategies, and it is often the case that one or more intermediate rasters must be computed and stored from a tedious succession of functional operations. With proper training however it is envisaged that better results could be obtained from a neural network solution that incorporated either standard inputs or standard inputs plus additional terrain-based inputs; the latter facilitating a more informative description of the local point-based area. Such models could also incorporate two or more simple processing operations and generate appropriate data values "on-the-fly" - thus reducing the overall intermediate data storage requirements.

Morrison's trigonometric surface (Morrison, 1971; 1974) is a single equation that takes the form of 49 sine and cosine terms all added together, and represents a least-squares fit to 121 data points read from a square lattice on Hsu & Robinson's (1970) Surface III, which is a real topographic map. This "equation surface" can be processed using the symbolic manipulation methods of differential calculus to obtain a partial derivative of the original equation in both x and y directions (Jones, 1996). The true slope value at a particular point on the surface can then be determined from the partial derivatives in the manner described by Sharpnack & Akin (1969) i.e. gradient in the down dip direction. 5,000 random point samples were generated from the original equation and its partial derivatives, comprising a grid of nine elevation values with a 10 unit (100m) offset, together with local slope value for the central point - the co-ordinates for this operation being held in a random lookup table generated from a uniform distribution. All five variables were then subjected to linear normalisation between zero (lowest possible value for that variable in the dataset) and one (highest possible value for that variable in the dataset). The normalised file was split into two equal data sets; one for training the network, the other for split-sample validation purposes.

The Stuttgart Neural Network Simulator was used to construct a two-hidden-layer feedforward network with a 9:12:12:1 configuration and with all appropriate connections enforced. The input nodes were for the nine elevation values. The output node was for true central slope.

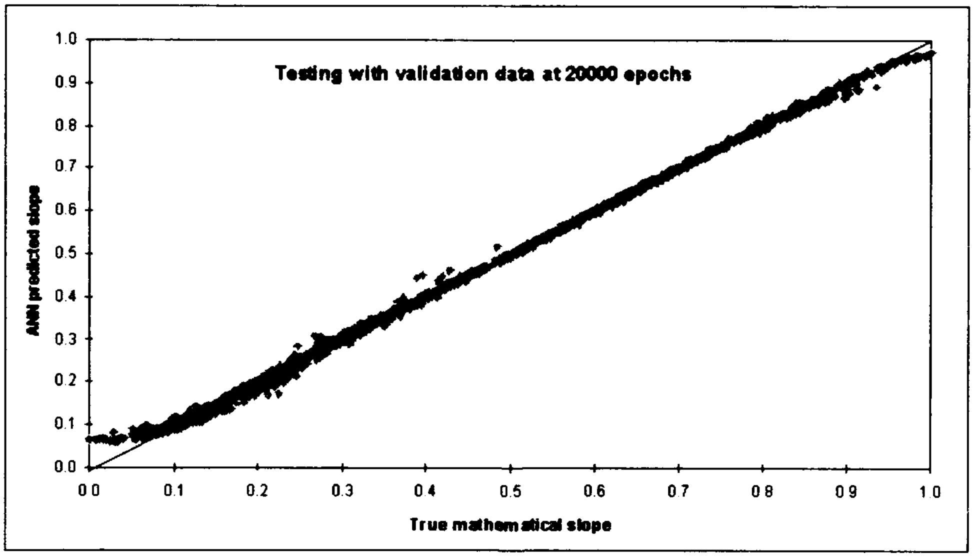


Figure 6: Predicting slope values from grid squares on a trigonometric surface with an artificial neural network. Scale is in normalised units. Line of perfect agreement drawn in black.

Network training was undertaken using "backpropagation without momentum". The learning rate was reduced according to a sliding scale at pre-set intervals and the network was observed to converge in a somewhat irregular manner - with the error curve following a jagged, staircased, downhill path - thus suggesting this might not be the most efficacious modelling solution. Training was stopped at 20,000 epochs. Error reduction was observed to be almost non-existent at this point indicating broad-scale convergence. The final sum squared error for the output node was just over 0.06 normalised units. At each change in the learning rate both training and validation datasets were passed through the trained network in its non-training mode and network output plotted against model output. In all instances the two plots were quite similar - which is indicative of good generalisation and modelling. A scatterplot for the validation data at the final stage of the learning process is reproduced here in Figure 6. Throughout most of the plotting space the output data exhibits a limited spread of values, and although there are one or two minor outliers in the central region, the general trend has a close association with the line of perfect agreement - with values both above and below it - thus providing further evidence of good modelling. However, notable discrepancies exist at both upper and lower ends of the graph, where the network has failed to predict correct results. Whether these problems are related to architectural considerations or inadequate deterministic inputs is a matter for further investigation. This experiment altogether demonstrates the unharnessed potential of using neural networks to process map based data at both individual cell and localised grid square levels.

5. Conclusions

Various map-based modelling tasks have been attempted and good results were achieved. So, given that in all instances little or no effort was made to achieve an optimal solution, for example in terms of different network architectures or data input formats - the inference from these experiments must be that real possibilities do exist for using neurocomputing solutions to perform geographical information storage and processing. More rigorous detailed experimentation should therefore be undertaken in order to advance the current state of knowledge in this area of science.

6. Acknowledgements

The Stuttgart Neural Network Simulator (SNNS) was developed in the Institute for Parallel and Distributed High Performance Systems at the University of Stuttgart[1].

Jo Cheesman, Manchester Metropolitan University, collaborated in experiment two.

Kevin Jones, Macaulay Land Use Research Institute, provided the equations for experiment three.

7. References

Brakensiek D.L. and Rawls W.J. (1983) Use of infiltration procedures for estimating runoff, *paper presented to the Soil Conservation Service National Engineering Workshop, Tempe, Arizona.*

Brooks R.H. and Corey A.T. (1964) Hydraulic properties of porous media, *Hydrology Paper No.3, Department of Civil Engineering, Colorado State University, Fort Collins, Colorado.*

Green N.H. and Ampt C.A. (1911) Flow of air and water through soils, *Journal of Agricultural Science*, 4, 1-24.

Hsu M. and Robinson A.H. (1970) *The Fidelity of Isopleth Maps*. Minneapolis: University of Minnesota Press.

Jones K.H. (1996) A comparison of eight algorithms used to compute slopes as a local property of the DEM, *Proceedings Fourth National Conference on GIS Research UK*, University of Kent at Canterbury, Canterbury, Kent 10-12 April 1996, 7-12.

Mitasova H. and Mitas L. (1993a) Interpolation by Regularised Spline with Tension: I. Theory and Implementation, *Mathematical Geology*, 25, 6, 641-655.

[1] URL = http://www-ra.informatik.uni-tuebingen.de/SNNSinfo/

Mitasova H. and Mitas L. (1993b) Interpolation by Regularised Spline with Tension: II. Application to Terrain Modelling and Surface Geometry Analysis, *Mathematical Geology*, 25, 6, 657-669.

Morrison J.L. (1971) Method-Produced Error in Isarithmic Mapping, *Technical Monograph No. CA-5, American Congress on Surveying and Mapping, Washington, DC.*

Morrison J.L. (1974) Observed Statistical Trends in various Interpolation Algorithms Useful for First Stage Interpolation, *The Canadian Cartographer*, 11, 2, 142-159.

Sharpnack D.A. and Akin G. (1969) An Algorithm for Computing Slope and Aspect from Elevations, *Photogrammetric Engineering*, 35, 3, 247-248.

Zhou Q. Wang P. and Pilesjö P. (1997) On the Quantitative Measurements of Errors Generated from Hydrological Modelling Algorithms, *Proceedings GIS AM/FM ASIA'97 & GeoInformatics'97*, Taipei, Taiwan, 26-29 May, 811-819.

Overcoming the Restrictions of Current Euclidean-Based Geospatial Data Models With a Set-Oriented Geometry

Jochen Albrecht
Department of Geography
The University of Auckland
j.albrecht@auckland.ac.nz

Presented at the second annual conference of GeoComputation '97 & SIRC '97, University of Otago, New Zealand, 26–29 August 1997

Abstract

This paper advocates the *spatialisation* of common data analysis by applying the techniques used in (geo-)spatial data analysis to any form of attribute space. It proves the distinction between spatial and non-spatial data to be an artificial one. New light is thrown onto the discussion of the relational database model and its applicability in spatial procedures. At the same time, the somewhat halted investigation of simplicial structures as a means of analysing spatial relations is expanded by using them as a form of representation of combinatorial concepts.

1 Introduction

Classification schemes are an impress of the human brain on a set of data (Gould, 1981, p. 299), they do not depict reality and destroy the richness of ambiguity. Research in combinatorial mathematics and algebraic topology (Atkin, 1974) and the number crunching capabilities of today's computers help us to handle *creatively* the additional complexity that set-based approaches imply. All the current research in fuzzy technologies (Davis and Keller, 1996; Dawson and Jones, 1995; Jiang and Kainz, 1996; Molenaar, 1996; Usery, 1996) bear witness of the renewed recognition of the richness of ambiguity as the backcloth that holds our data together.

One of the biggest myths in spatial analysis is the singularity of spatial data. While there is some justification to the fact that, historically, the efficient access of large spatial databases required special data structures (Samet, 1990), this should hardly be a handicap anymore (Varma, 1991; Nieuwenhuijs, 1995), and it is time that we free ourselves of the mental straight jacket that current geographic information systems (as well as other geospatial software) impose on us.

Space is one of the fundamental human experiences. Cognitive studies (Mark, 1989; Golledge, 1990; Nyerges, 1992) prove that people tend to "spatialise" many aspects of their everyday life. As such, spatial metaphors (Kuhn, 1992) are powerful means of categorisation (Rosch and Lloyd, 1978; Lakoff and Johnson, 1980; Johnson, 1987; Lakoff, 1987), they help us to structure the complexity of reality. Research in multidimensional domains, such as in environmental applications, face a similar problem of complexity. However, they do not yet employ such cogent concepts like neighbourhood, proximity, or shape. The methodology introduced in this paper overcomes the schism between spatial and non-spatial data by treating each non-spatial category as another dimension. Since the analysis of high-dimensional data is difficult to conceptualise, methods of combinatorial topology will be used to represent and reason in n-dimensional space. This paper is the third in a series of publications (Albrecht and Kemppainen, 1996; Kemppainen and Albrecht, 1996) where the author develops a formal framework for the extensibility of spatial operators and the first where algebraic specifications as well as graphical language are employed to overcome the limitations of current representations of the spatial domain.

2 Spatialisation of Attribute Space

The traditional concepts of spatial reasoning do not need to be restricted to the two geometrically defined variables that are usually employed to describe a given location on the earth's surface. They can rather be applied to any conceivable space and thus create *attribute landscapes*. Following Goodchild's (1990) definition, geographic information consists of a location x, y and parameters $<z_1, z_2, ..., z_n>$ measured at this location forming a tuple $T = <x, y, z_1, z_2, ..., z_n>$. In general, these variables can be mapped to a continuous scale and represented as vectors, or more specifically, as axes that put up an *n*-dimensional space. Thus, each attribute can be a dimension just as time is used as one dimension in change detection analysis (Macleod et *al.*, 1993; Yuan, 1996), although it will be argued further down that time is *n*-dimensional as well. Since each axis is a vector, all the topology-based rules of spatial reasoning apply in these non-geometric, better: not earth-referenced, spaces. Hence the term 'spatialisation' of attribute space.

The spatial metaphor may just as well be employed for attribute landscapes and the mere existence of scatter plots, as they can be found in a number of current statistical packages, proves that this idea is not far-fetched. However, this scarcely utilises the abundance of spatial metaphors for explorative data analysis (Aspinall and Lees, 1994). One example for innovative use of attribute space in a geospatial application is Gahegan (1996) who visualises attribute landscapes using slope, flow length and accumulation to depict a hydrological data space.

One way to represent those *n*-dimensional spaces is by depicting each *n*-dimensional object as a *n*-simplex (see *Table 1* and the corresponding *Figures 1-3* in the appendix).

The concept of Q-analysis[1] introduced by Atkin (1974, 1977, 1981) provides a uniform formal, set-based approach to the definition of space that lends itself to the definition of a new spatial data model. Attempts into this direction have been made by Egenhofer and Herring (1990), however, their approach is geared toward geometry only and lacks the degrees of freedom offered by Q-analysis. Due to length constraints enforced by the editors, only the additional advantages of Q-analysis over what has already been presented in the Maine school will be presented here. The interested reader is referred to Vanacek and Ferrucci (1991) and Faltings (1995).

Each row in *Table 1* can be represented by a simplex (see *Figure 3 a, b, c, f, g*). Together they form a simplicial complex KY(X, I) where I symbolises the relation between X and Y (the whole process could be inverted by looking down the columns and thereby analysing KX(Y, I)). Each simplex can be dissected into the faces that it is made of. The easiest example is the three-dimensional Y_5 (see *Table 2*). It consists of one 3-dimensional face (a tetrahedron), four 2-dimensional faces (triangles), six 1-dimensional faces (lines) and four 0-dimensional faces (points). The faces are *q*-connected if they share (*q* + 1) vertices. The notion of *q*-level can be understood as a kind of filter that restricts the

[1] Atkin's Q-analysis is an unfortunate denominator for an analytical technique that has nothing whatsoever to do with Q-mode analysis known in factor analysis.

λ	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9	X_{10}
Y_1	0	1	1	0	1	0	1	0	0	1
Y_2	1	0	1	1	0	1	1	1	0	0
Y_3	0	1	0	0	1	0	0	1	1	1
Y_4	1	1	0	1	1	0	1	1	0	0
Y_5	1	0	0	0	1	1	0	1	0	0

Table 1 A simplicial complex in matrix form.

q-level	$Y_5(X, \lambda)$
3	$\{X_1\ X_5\ X_6\ X_8\}$
2	$\{X_5\ X_6\ X_8\}$ $\{X_1\ X_6\ X_8\}$ $\{X_1\ X_5\ X_8\}$ $\{X_1\ X_5\ X_6\}$
1	$\{X_1\ X_5\}$ $\{X_1\ X_6\}$ $\{X_1\ X_8\}$ $\{X_5\ X_6\}$ $\{X_5\ X_8\}$ $\{X_6\ X_8\}$
0	$\{X_1\}$ $\{X_5\}$ $\{X_6\}$ $\{X_8\}$

Table 2 Dissection of the n-dimensional simplex Y into 0-(n-1)-dimensional faces

view onto the simplex to the dimensionality q. This filter has the same effect as living in Abbot's (1884) 'Flatland' where it is impossible for its inhabitants to perceive anything of higher dimensionality (with all the consequences marvellously described in this little novel).

The dimensionality of a simplex is called top-q ($\hat{q}$) and the dimensionality at which it begins to connect with other simplices bottom-q ($\check{q}$). These two indicators can be employed to define a measure of

Eccentricity $$e = \frac{(\hat{q}+1)-(\check{q}-1)}{(\hat{q}+1)}$$

which neatly describes the connectivity of a particular simplex in comparison to global connectivity within the simplicial complex. A global measure of structure is given by the *structure vector* Q, which is determined by the number of q-connected components for values of q from 0 to dim K. The simplicial complex in *Figure 1* is fairly well interconnected allowing for high-dimensional traffic between the individual simplices. *Figure 4* is a realistic example of an organisation consisting of several well-functioning departments which are only loosely connected via 1-dimensional simplices. Communication across the q-hole is restricted to this one dimension and the effect is the same as the experience of a cube living in Flatland. Depending on the degree to which a q-hole reduces connectivity within the complex, it can be regarded as an obstructive object of value *q*. Applied to social activities, q-holes describe the limitation of freedom that an individual experiences in a particular structure.

3 Relevance to the Application of GIS Analysis in Social Sciences

Couclelis and Gale (1988) define a hierarchy of higher-level spaces that start with the conventional Euclidean space but then extends to physical, sensorimotor, perceptual, cognitive and finally symbolic space. While they focus on algebraic group theoretical notions, it is presumed here that these spaces can only be sustained by releasing some of the restrictions of higher-level mathematical spaces, i.e., advancing the hierarchy of Couclelis' spaces we have to descend the ladder of mathematical spaces. Arguably, the best setting for reasoning in cognitive spaces is then relational space with some of the methods of Q-analysis.

Cognition is based on experience. We constantly enhance *our cognition by adding experiences to our structural* model of the world with our current cognitive state being a cover set of all previous experiences (faces) of what constitutes the (simplicial) complex of our cognitive world model. Q-analysis is the only tool known to the author that allows to analyse the natural complexity of experiential space-time, especially with respect to the parallel universes of members of a community. It opens a whole new set of opportunities for the analysis of time lines (/prisms) introduced by Hägerstrand (1975), and only sporadically followed up (e.g. Miller, 1991; Forer, 1993). As an add-on, it fuels the discussion about the nature of time as discussed in various standardisation committees (e.g. the International Standards Organisation's technical committees on Geoinformation or Structured Query Languages) who ponder about its characterisation by attributes or as *a* dimension. *Figure 5* depicts the at least two-dimensional nature of time.

The comparison of different structure vectors and the analysis of obstruction areas provides insight to the forces that result in a particular structure (and if the matrix includes locational information to the social construction of space (Gregory and Urry, 1985; Lefebvre, 1991)). Q-analysis is a tool that permits the simulation of different scenarios for changes in the (infra-) structure of data matrix.

4 Conclusion

In earlier papers (Albrecht and Kemppainen, 1996; Kemppainen and Albrecht, 1996) the author argued that there (1) exists a hierarchy of spaces, and (2) the GIS domain would profit from the definition of spatial operations at as low a conceptual space as possible. With this paper it could be shown that we lose little (there is some ambiguity introduced by working in less constrained spaces) but gain a lot in additional analytical power by employing the richness of data that has not been normalised (for use in relational databases) nor restricted to strict functional relationships which hide higher-dimensional patterns of whatever data set we examine. This way, several birds can be killed with one stone. The unpleasant dichotomy between spatial and non-spatial analysis can be resolved and for the first time we have the opportunity to overcome the schism between what is accused to be a positivist tool and adherents of post-structural scientists in the humanities (Pickles, 1995).

5 References

Abbot, E. (1884) *Flatland: A Romance of Many Dimensions.* Available as http://www.tiac.net/users/eldred/eaa/FL.HTM

Albrecht, J. and H. Kemppainen (1996) A framework for defining the new ISO standard for spatial operators, *Proceedings 4th National Conference GIS Research 1996 (GISRUK'96)*, Canterbury, pp. 78-84.

Aspinall, R. and B. Lees (1994) Sampling and analysis of spatial environmental data, *Proceedings 6th International Symposium on Spatial Data Handling*, pp 1086-1098.

Atkin, R. (1974) *Mathematical Structure in Human Affairs*, London: Heinemann.

Atkin, R. (1977) *Combinatorial Connectivities in Social Systems: an application of simplicial complex structures to the study of large organizations.* Basel: Birkhäuser.

Atkin, R. (1981) *Multidimensional Man*, London: Penguin.

Benwell, G (1993) On the Spatial Analysis of Hieracium and Towards a Fuzzy Representation. *Proceedings of the 5th Annual Spatial Information Research Centre.* **1**: 279-288, Dunedin: University of Otago, New Zealand..

Couclelis, H. and N. Gale (1988) Space and Spaces. *Geografiska Annaler* **68B**:1-12.

Davis, T. and P. Keller (1996) Modeling and Visualizing Uncertainties in Environmental Assessment Using Fuzzy Logic and Monte Carlo Simulation. In Heit, Parker and Shortreid (Eds.) *GIS Applications in Natural Resources* pp. 125-132. Fort Collins: GIS World, Inc.

Dawson, T. and Ch. Jones (1995) Representing and Analyzing Fuzzy Natural Features in Geographical Information Systems. In *GIS '95 Conference Proceedings.* 1:423-428. Fort Collins: GIS World, Inc.

Egenhofer, M. and J. Herring (1990) A mathematical framework for the definition of topological relationships, *Proceedings of the 4th International Symposium on Spatial Data Handling*, Zürich, pp 803-313.

Faltings, B. (1995) Qualitative Spatial Reasoning using Algebraic Topology. *Proceedings COSIT '95: Spatial Information Theory: A Theoretical Basis for GIS*, pp. 17-30. Lecture Notes in Computer Science. Berlin: Springer-Verlag.

Forer, P. (1993) An implementation of a discrete 3-D space-time model of urban accessibility. *Proceedings 17th New Zealand Geographical Society Conference*, pp. 51-57.

Gahegan, M. (1996) Visualisation strategies for exploratory spatial analysis, *Proceedings Third International Conference/Workshop on Integrating GIS and Environmental Modeling*. An electronic version is available as

http://www.ncgia.ucsb.edu/conf/sf_papers/gahegan_mark/gahegan.html.

Golledge, R. (1990) *Some relations between geographical and psychological space: the case of configurational (survey) knowledge.* Paper presented at the Boston Meeting of the American Psychological Association, August 1990. Reported in Nyerges 1992.

Goodchild, M. (1990) Geographical Data Modeling, in Frank, A. and M. Goodchild (eds.) *Two Perspectives on Geographical Data Modeling*, NCGIA Technical Report

90-11, Santa Barbara: National Center for Geographic Information and Analysis.

Gould, P. (1981) A s: uctural language of relations, in Craig, R. and M. Labovitz (Eds.) *Future Trends in Geomathematics*, London: Pion.

Gregory, D. and J. Urry (1985) *Social Relations and Spatial Structures*. Basingstoke: Macmillan.

Hägerstrand, T. (1975) Space, time and human conditions. In Karlqvist, Lundqvist, L. and F. Snickars (Eds) *Dynamic Allocation of Urban Space*, pp. 3-14. Farnborough: Saxon House.

Jiang, B. and W. Kainz (1996) Fuzzy overlay analysis with linguistic degree terms. In *Advances in GIS Research II: Proceedings 7th International Symposium on Spatial Data Handling*. 1:5B.25-42. Delft: Faculty of Geodetic Engineering, Delft University of Technology.

Johnson, R. (1987) *The Body in the Mind: the bodily basis of meaning, imagination, and reason*, Chicago: The University of Chicago Press.

Kemppainen, H. and J. Albrecht (1996) Scalability of Spatial Operators, *Proceedings GIS/LIS'96*, pp. 162-172. Denver.

Kuhn, W. (1992) Paradigms of GIS Use. *Proceedings 5th International Symposium on Spatial Data Handling*, pp 91-103.

Lakoff, G. and M. Johnson, M. (1980) *Metaphors We Live By*, Chicago: The University of Chicago Press.

Lefebvre, H (1991) *The Production of Space*. Translation, original from 1974. Oxford: Blackwell.

Macleod, R., Congalton, R. and F. Short (1995) Using Quantitative Accuracy Assessment Techniques to Compare Various Change Detection Algorithms for Monitoring Eelgrass Distributions in Great Bay, N.H. Generated from Landsat TM Data. *ACSM/ASPRS Annual Convention and Exposition Technical Papers*, 3:876-885. Bethesda: ACSM/ASPRS.

Mark, D. (1989) Cognitive image schemata for geographic information: relations to user views and GIS interfaces. *Proceedings GIS/LIS'89*, . Falls Church: ASPRS/ACSM, pp 551-560.

Miller, H. (1991) Modelling accessibility using space-time prism concepts within geographical information systems. *International Journal of Geographical Information Systems*, 5(3):287-301.

Molenaar, M. (1996) A Syntactic Approach for Handling the Semantics of Fuzzy Spatial Objects, in Burrough, P. and A. Frank (Eds.) *Geographic Objects with Indeterminate Boundaries*. pp. 207-224. Bristol, PA: Taylor & Francis Inc.

Nieuwenhuijs, S. (1995) Oracle MultiDimension: new frontiers in spatial data management, *GIS Europe*, 6, pp 40-42.

Nyerges, T. (1992) Thinking with geographic information: user interface requirements for GIS, in Mark, D. and A. Frank (eds.) *User Interfaces for Geographic Information Systems*, NCGIA Technical Report 92-3, Santa Barbara: National Center for Geographic Information and Analysis.

Pickles, J. (Ed.) (1995) *Ground Truth: The Social Implications of Geographic Information Systems*. New York: The Guilford Press.

Rosch, E. and B. Lloyd (eds.) (1978) *Cognition and Categorization*, Hillsdale: Erlbaum.

Usery, L. (1996) A Conceptual Framework and Fuzzy Set Implementation for Geographic Features, in Burrough, P. and A. Frank (Eds.) *Geographic Objects with Indeterminate Boundaries*. pp. 71-86. Bristol, PA: Taylor & Francis Inc.

Vanacek, G. and V. Ferrucci (1991) A Spatial Index for Convex Simplicial Complexes in d Dimensions. *Second Symposium, SSD '91: Advances in Spatial Databases*, pp. 361-380. Lecture Notes in Computer Science 525, Zürich: Springer-Verlag.

Varma, H. (1991) *Implementation of topological relationships using SQL* (no place given).

Yuan, M. (1996) Modeling semantical, temporal, and spatial information in geographic information systems, in Craglia, M. (Ed.) *Progress in Trans-Atlantic Geographic Information Research*, First ESF-GISDATA and NSF-NCGIA Summer Institute on Geographic Information, Bristol: Taylor and Francis.

Appendix

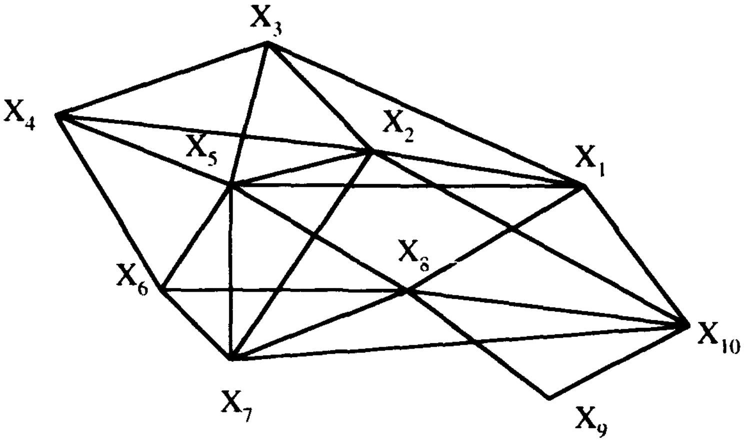


Figure 1 Simplicial complex KY(X, l) representing the matrix of Table 1.

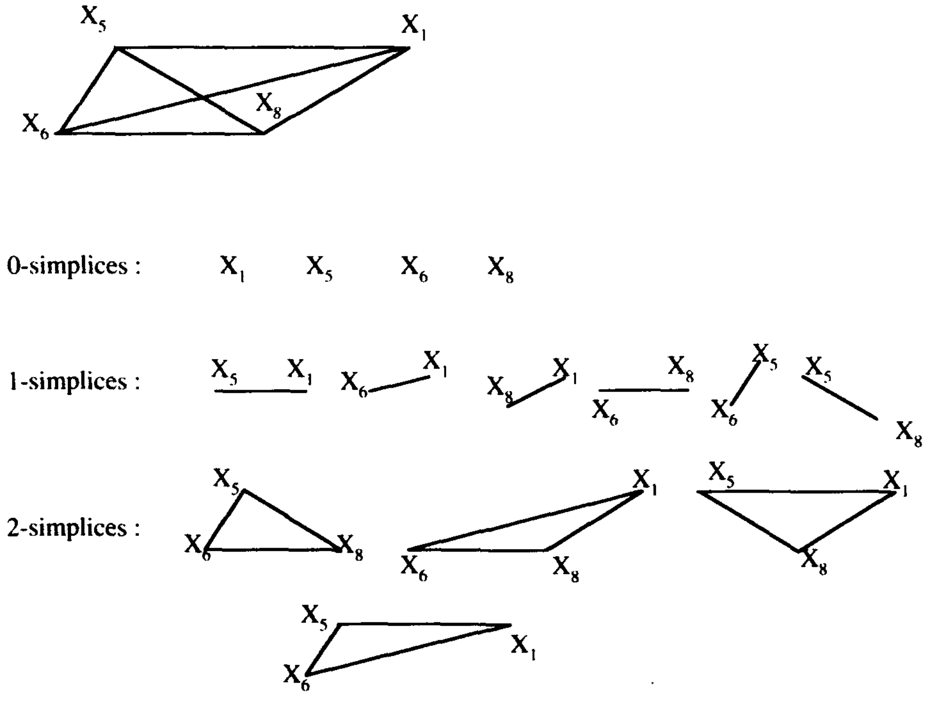


Figure 2 The 3-simplex Y5(X) and its decomposition in 0-, 1-, and 2-dimensional faces

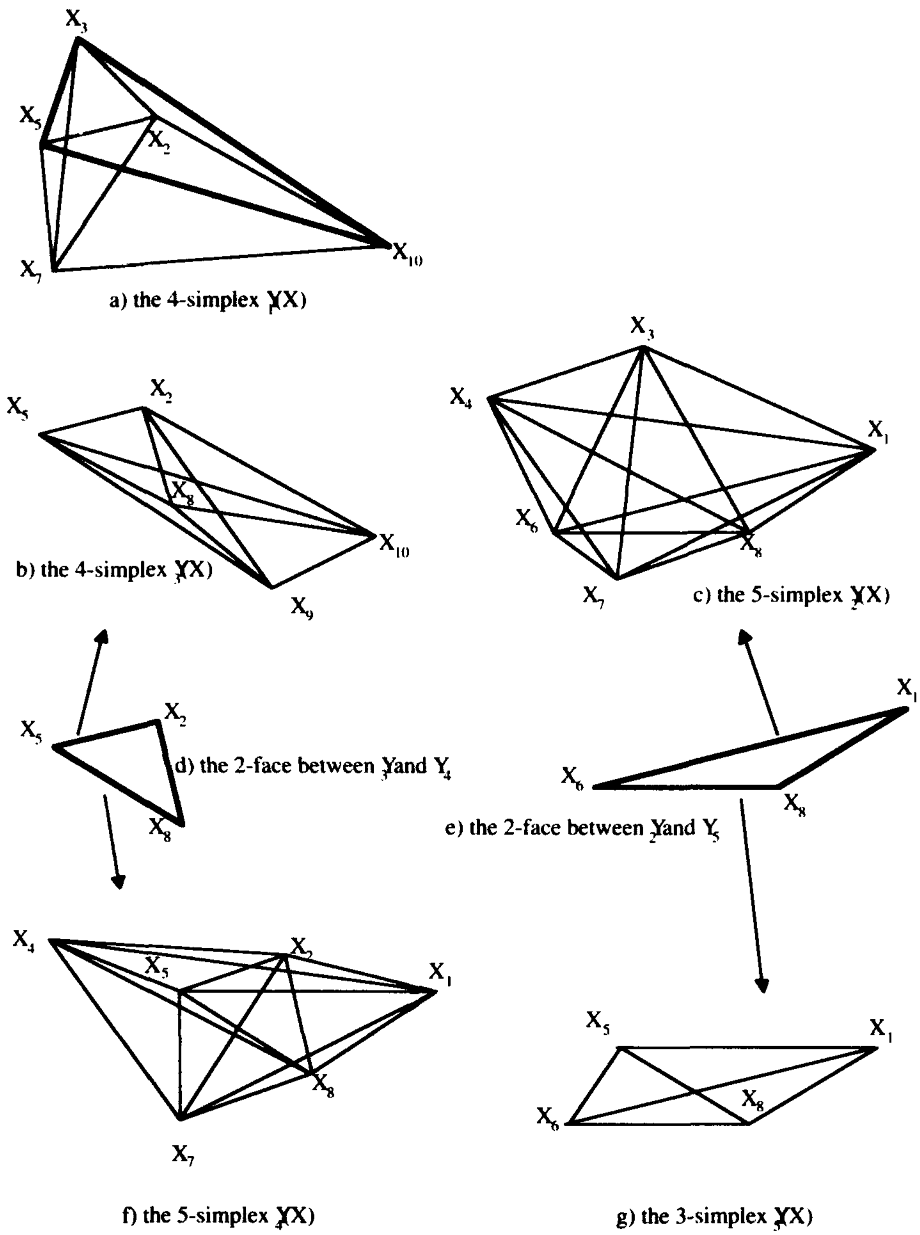


Figure 3 The five simplices that make up KY(X) with two examples of n-dimensional (communication inter-) faces between two members of KY(X) each.

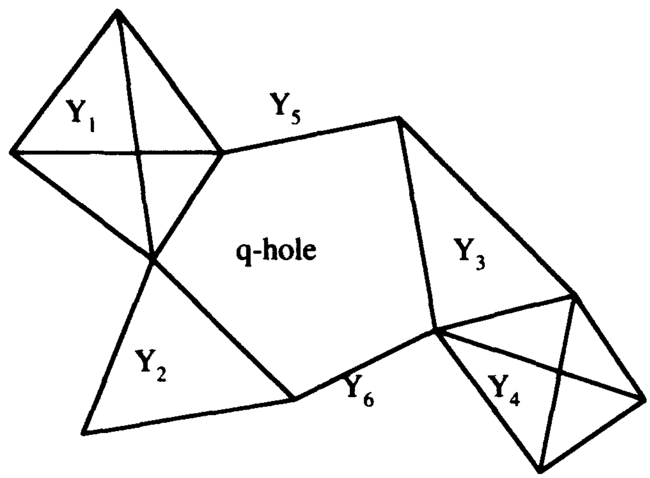


Figure 4 q-hole narrowing 2-dimensional traffic down to 1-dimensional one.

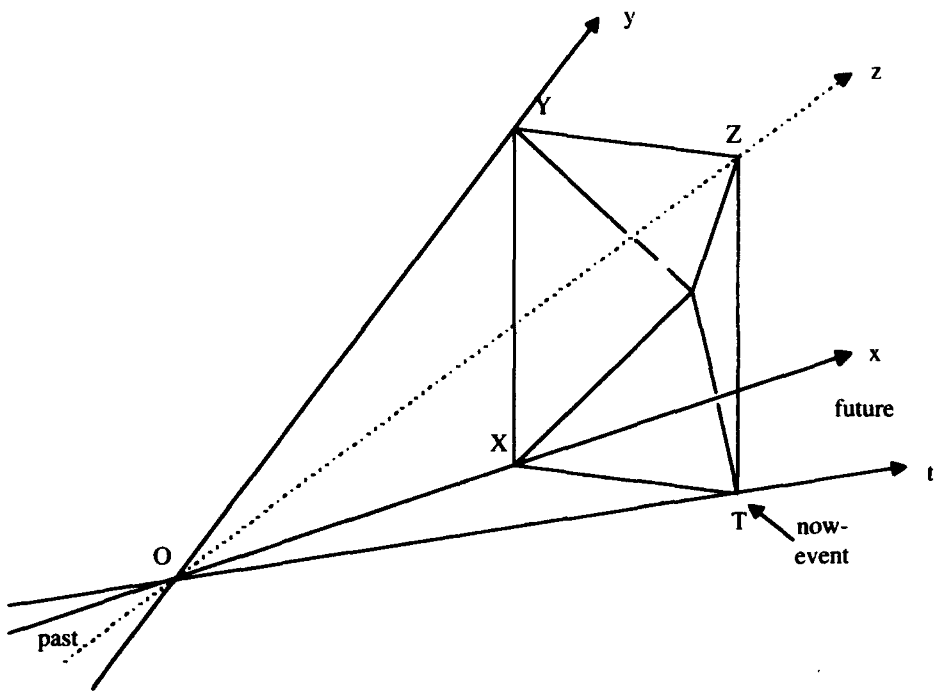


Figure 5 Five-dimensional space-time continuum (the 2-dimensional shaded time-interval must be thought to be expanded to the other Euclidean dimensions as well) (adapted from Atkin 1981).

Meso-scale mapping of soil temperatures in the Mackenzie Basin, New Zealand

James R. F. Barringer
Landcare Research NZ Ltd.,
P.O. Box 69, Lincoln 8152,
Canterbury, NEW ZEALAND

Presented at the second annual conference of GeoComputation '97 & SIRC '97, University of Otago, New Zealand, 26-29 August 1997

1. INTRODUCTION

In mid-latitude areas, variation in local temperatures associated with topography is a significant factor that needs to be accommodated in environmental models and future land-use strategies. Qualitatively, variation of temperature within complex landscapes is well understood. However, quantification of these patterns has been limited. Temperatures measured at standard climate stations give a broad indication of spatial and temporal variations in regional climate but do not explain local patterns of climate variation. However, sites for climate stations are selected to conform to standard conditions (i.e., flat, large fetch, unshaded, etc.) to allow easy comparison between stations. Data collected at such sites are frequently not representative of much of the surrounding area, and estimating local climate from nearby standard sites is likely to involve significant error.

Previous attempts to explain spatial variation of temperature in New Zealand have focussed largely on empirical modelling and/or spatial interpolation from existing standard climate station data. Typically, site properties such as latitude, altitude and distance from coast are utilised in these approaches (Norton, 1985). While these may provide acceptable regional results, they are not adequate for establishing local scale variability where factors such as aspect are significant. Of the meso-scale climate studies in New Zealand only Turner and Fitzharris (1986) have explicitly sampled a landscape. However, their sample did not incorporate factors which could easily be mapped automatically (e.g., elevation, slope, or aspect). This makes it difficult to interpolate local climate to other adjacent areas. This paper describes the development and testing of an empirical model for predicting the spatial variability of soil temperature within the central South Island high country of New Zealand which is based on site characteristics which can be easily derived from a digital elevation model (DEM).

2. METHOD

Temperature data collection

Temperature data were collected from 43 sites in the Grampians range in a stratified sample grouped by elevation and aspect (Fig. 1). Sites ranged in altitude from 600 to 1800 metres, and within each altitudinal stratum sites approximating the four primary aspects (i.e., north, east, south, and west) were sampled. At Glentanner some 50 km to the north west, a further 27 sites provide a similarly stratified and replicated sample from 800 m to 1400 m.

Attributes recorded for each site included altitude, aspect, and slope. Aspect was recoded into degrees south of north (i.e., 0 - 180°). Temperature data were collected quarterly (February, May, August, November) using a hand held digital thermometer probe lowered down 20 mm diameter PVC access tube embedded 0.75 m into the ground with its lower extremity sealed by a protruding aluminium alloy plug to provide good thermal contact with the soil. A water/anti-freeze mix in each tube (to a depth of approximately 5 cm) provided good thermal contact for the temperature probe.

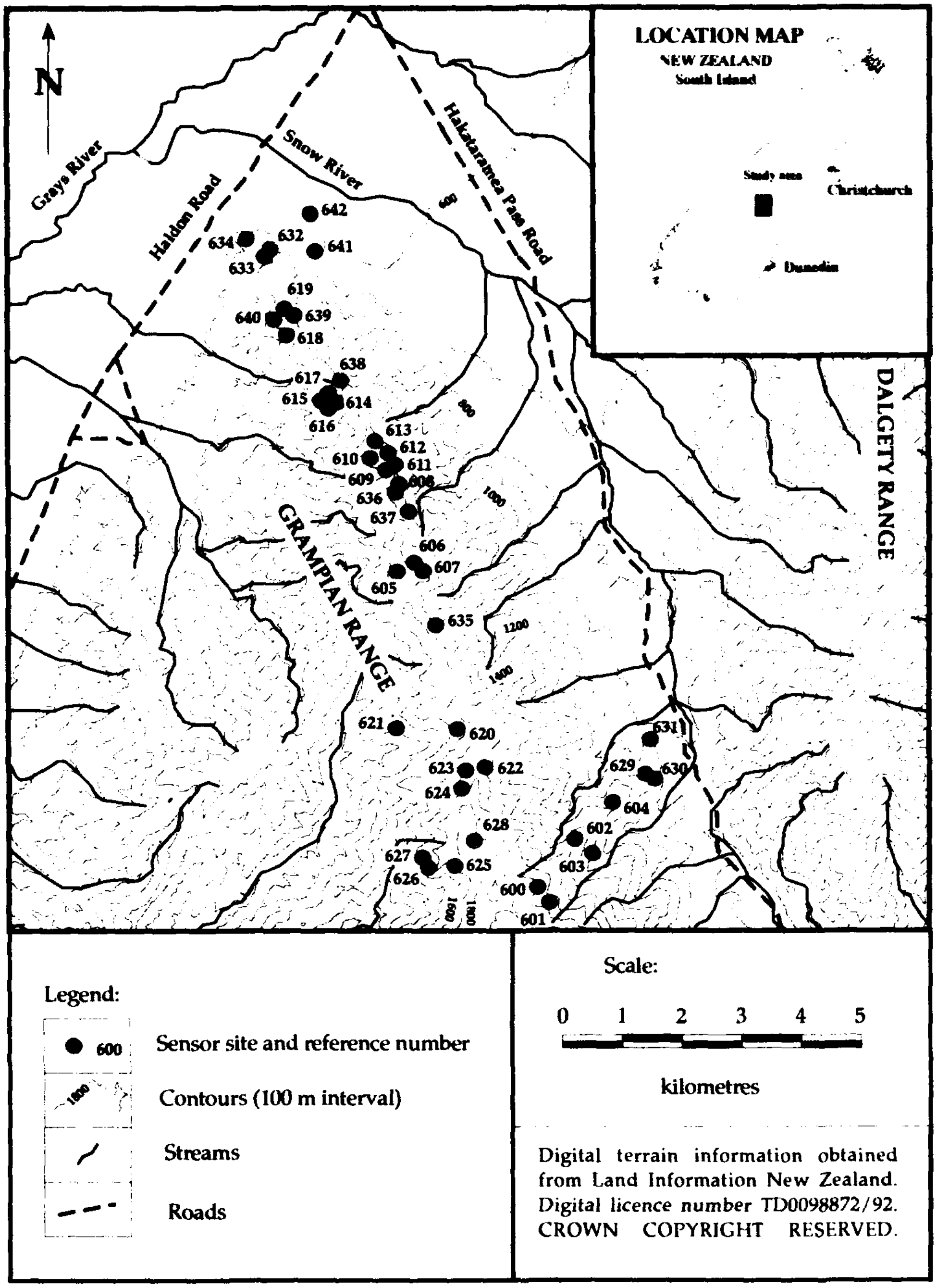


Figure 1　*Location map, and sampling scheme for the 43 sites in the Grampians Range.*

The site attribute data (independent variables) and soil temperature data (dependent variable) were analysed using multiple linear regression to derive an empirical model relating soil temperature to site characteristics (altitude, aspect and slope) for each season. Data from both networks of sites were used to calculate regression interaction terms to determine inter-site variability of regression coefficients, giving an indication of the model's applicability to the surrounding area, which is one where a significant rainfall gradient occurs.

3. RESULTS

Correlation coefficients (r^2) from the seasonal regression analyses were high, ranging from 0.83 to 0.96. These seasonal regression models, a 25 metre resolution DTM, and raster-based GIS were used to derive maps illustrating patterns of soil temperature variation over the 144 km^2 study area (e.g., Fig. 2).

Regression interaction terms between elevation and location (0.001), and aspect and location (-0.003 to 0.001) are small (Table 1), and clearly indicate that the regression

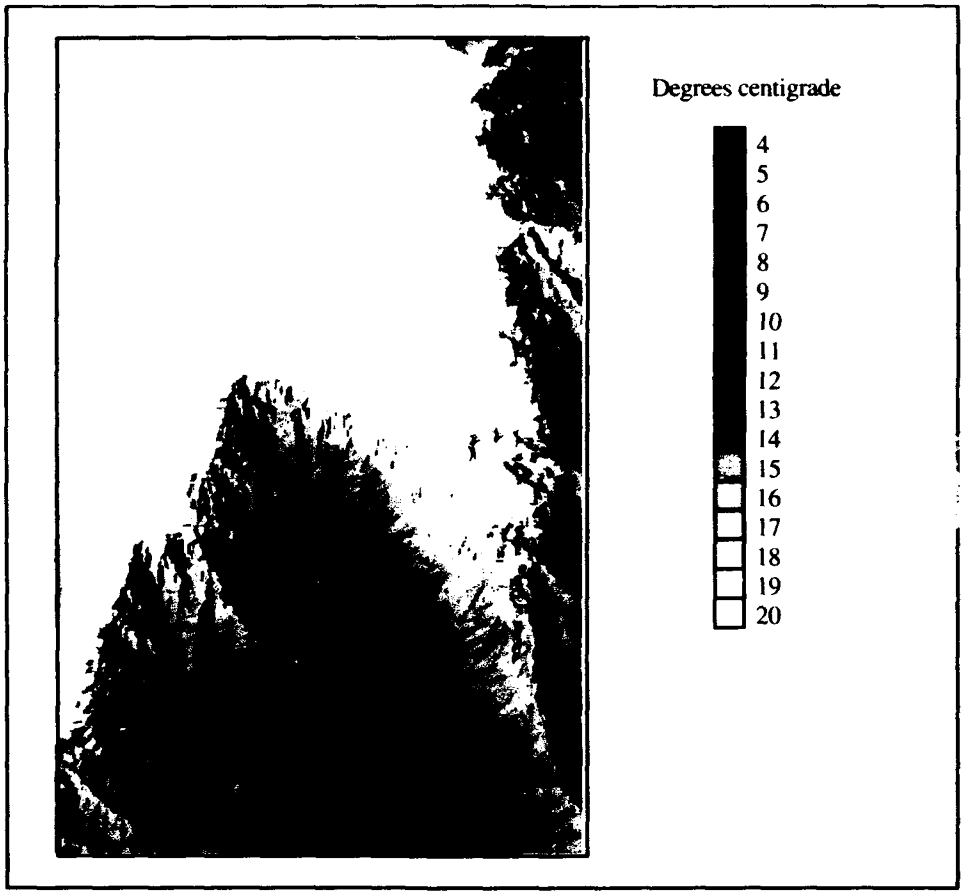


Figure 2 Map of soil temperatures for the Grampians Range, February (summer) 1992, based on regression model and 25 m resolution DEM: SoilTemp = 23.188 - 0.007 (altitude) - 0.022 (aspect) - 0.016 (slope), r^2 = 0.893.

Date	Constant	Altitude	Aspect	Slope	Location/ altitude	Location/ aspect	Location/ slope
February 93	20.897	-0.007	-0.018	0.014	0.001	0.001	-0.028
May 93	14.694	-0.006	-0.025	0.039	0.001	-0.003	-0.026
August	939.161	-0.005	-0.025	0.051	0.001	0.000	-0.039
November 93	18.028	-0.008	-0.022	0.030	0.001	-0.002	-0.019
1993	15.692	-0.007	-0.022	0.033	0.001	-0.001	-0.028
February 94	23.391	-0.007	-0.024	0.031	0.001	0.001	-0.017
May 94	14.903	-0.006	-0.029	0.048	0.001	-0.003	-0.022
August 94	8.257	-0.005	-0.015	0.023	0.001	-0.002	-0.035
November 94	18.365	-0.009	-0.018	+0.020	0.001	-0.003	-0.004
1994	15.837	-0.006	-0.022	+0.021	0.001	-0.002	-0.025

Table 1: *Results of combined regression analysis. The constant, altitude and aspect columns represent regression coefficients for data from both sites combined. The locational interaction terms give an indication of the difference in regression coefficient values between the two sites.*

model is sufficiently robust to be applied throughout the Mackenzie Basin (an area in excess of 15000 km^2).

4. CONCLUSIONS

The results of this survey provide a good picture of the patterns of spatial variation in soil temperature. The strikingly stable relationship between temperature variation and altitude, and to a lesser extent aspect, suggests that the regression coefficients determined from this study will be applicable over significant areas (at least 15%) of the South Island high country. Establishing a network of study sites in coastal areas and over a wider latitudinal range could yield more universally applicable models.

5. ACKNOWLEDGEMENTS

Land Information New Zealand (Digital Licence Number TD0098872/92). This project was funded by the Foundation for Research, Science and Technology under FRST contract C09325.

6. REFERENCES

Norton, D.A. (1985). A Multivariate Technique for Estimating New Zealand Temperature Normals. *Weather and Climate*. 5: 64 - 74.

Turner, A. and Fitzharris, B. B. (1986). Mapping Warm Season Degree Days at the Local Scale. *New Zealand Geographer*. 42: 57 - 64.

Issues in Representing Spatially Embedded (aphs

Malathy Naguleswaran
Department of Computer Science
University of Canterbury
Christchurch, New Zealand
Tel: (64) (03) 3518595
email: malathy@cosc.canterbury.ac.nz

Presented at the second annual conference of GeoComputation '97 & SIRC '97, University of Otago, New Zealand, 26-29 August 1997

1 Introduction

Graphs are an ubiquitous part of Computer Science, as well as many other fields. Spatially embedded graphs are a common feature of many urban geographic applications. Transport, utility, population all occur along spatial networks. The study of abstract graphs is an established field within discrete mathematics. The study of spatially embedded graphs however is a relatively new field. Commercial packages that handle spatially embedded graphs do so at the most basic level (ESRI 1992) We have put together the issues identified by others and the author on this subject. A comprehensive design for handling spatially embedded graphs in geographic applications will be a goal worth exploring. The issues identified are,

1. A graph model with associated query language
2. Multiple representation of graphs
3. Subgraph maintenance
4. Dynamic segmentation
5. Spatial indexing of linear spatial objects

The first three issues above can be dealt with as graph issues without reference to their geometric attributes. The last two arise purely from geometric considerations.

2 Graph model

In terms of the entity relationship model a graph must be represented as a many to many cyclic relationship. The graph semantics is captured in the cyclic relationship. Implementation of this cyclic relationship using traditional DBMS requires this many to many relationship to be broken down to a one to many relationships (Figure 1). Thus the graph semantics is lost in the final representation leading to inefficient handling of graph traversal. Within the object oriented paradigm it is appropriate to consider new models to explicitly capture the graph semantics. Guting (1994) explores a graph model where the many to many cyclic relationships are explicitly recorded in a database schema thus capturing the graph semantics. The definition of an edge class which is a many to many relation between objects of another class defines a graph. An ordered collection of edge objects that are serially connected defines a path over the graph. The path concept over a single graph does not add much value to the data model. However it is possible to define multiple edge classes in the database schema. In other words we have multiple graphs that could share common objects as their nodes. Thus a path can be defined over edges from many edge classes. A good example is public transport (Guting 1994). The physical network of routes, lines over which services exist, and the time schedules can all be represented as separate graphs with shared nodes across the graphs.

3 Multiple Representation

As in any geographic applications multiple representation of the same physical graph must be considered. With respect to spatially embedded graphs multiple representations arises from two requirements. One is when data

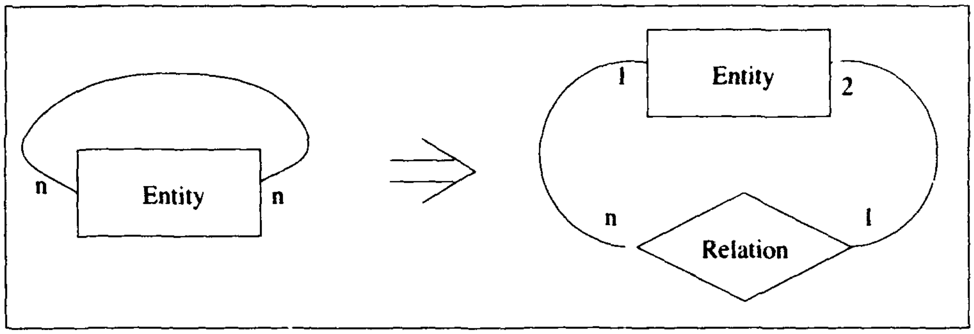


Figure 1

represented in disparate coordinate systems are to be treated as one seamless graph. Electrical networks with schematic diagrams of sub stations and cross section data of major junctions are an example. Newell et al (1994) provides a means of plugging these graphs together as a single graph for network analysis. The second use of multiple representation is to represent graphs at many scales and detail.

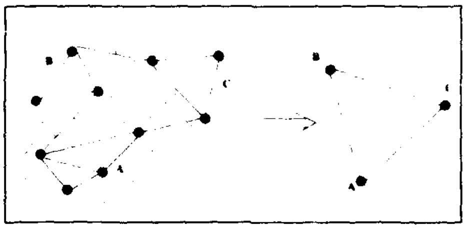


Figure 2

A simple solution to graph generalisation is shown in figure 2. Subgraphs(within the dotted line) can be treated as a node at a higher level with all edges leading out of the subgraphs to be treated as edges from this node at the higher level. Problems arise with this approach since the edges going out of the subgraph at the lower level may not adequately represent the edges from the node at the higher level. Therefore we also need to create edges at the higher level to represent those edges leading outside the subgraph.

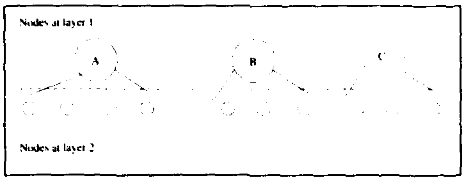


Figure 3

Mainguenaud (1995) outlines a solution which permits subgraphs to be represented in the above fashion. The solution is to setup the graph in layers where a strict hierarchy of nodes is maintained (Figure 3) Every node at a higher level is effectively a subgraph. Some of the nodes at a higher level may not expand into a subgraph at the next level but remain as a single node. Each layer has its o.vn collection of edges. The edges of a higher layer must also be related to the edges at a lower layer. The scheme has a strict and pre-wired hierarchy. Situations where different groupings down the layer is required exists. Location problems are a case in hand. These problems are solved using heuristic algorithm or linear programming techniques. In both cases grouping subgraphs into nodes can be used to reduce problem size. It is also known that different groupings can produce different results. Thus it is necessary to solve these problems with many different groupings to ensure that the obtained results are satisfactory.

4 Subgraph maintenance

Subgraphs are a recurring theme for handling spatially embedded graphs. We have already identified some of these in the previous sections. In a graph model the subgraph selection may be used to generate a graph reduced in size over which a graph query can be applied. Subgraph grouping for representing graphs at many levels of granularity was described in (section 3). We consider the theme of subgraph maintenance at a single abstraction level for a specific use.

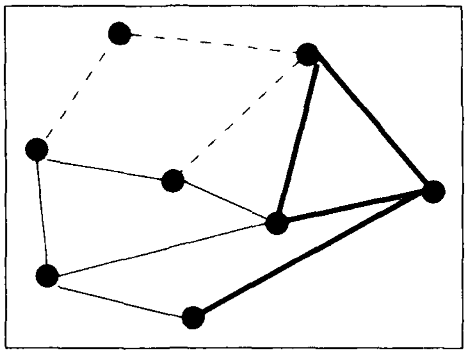


Figure 4

Location problems deal with the location and allocation of services among urban population. Population in an urban area can be modelled as a spatially embedded graph where the nodes and edges are loaded with population and services. The solution to the location problem is through heuristic graph algorithms to create subgraphs. The subgraphs thus generated must be maintained as the allocation zones for the services (Figure 4).

5 Access to geometry/Dynamic Segmentation

All issues of spatially embedded graphs considered so far have treated the graph edges at a certain abstraction level as a whole entity that is indivisible. At the physical level the graph edges of spatially embedded graphs have a linear geometry. Application requirements exists that need to access locations along this linear geometry. Examples are locating physical features along street

networks, address geocoding, urban zones represented as subgraphs where the subgraph includes only part edges of the base graph.

At the level of physical representation of spatially embedded graphs it is necessary to provide access to intermediate locations on the graph edges. Techniques used to provide this access have been referred to as dynamic segmentation (Dueker et al 1992). The object oriented paradigm can be exploited to provide a uniform interface for accessing linear geometry. Naguleswaran et al (1996) explores this idea further. Figure 5 depicts the idea. LSO (Linear Spatial Object) is an abstract class for the three types of linear geometry that can exist. BasicLSO is the representation of linear geometry using a series of points. SubLSO is a linear geometry defined as a section of an existing linear geometry. LinearConnectedLSO represents those geometry formed by the concatenation of many other linear geometries in a network.

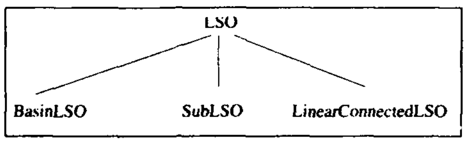


Figure 5

6 Spatial indexing of linear spatial ob jects

The family of PM trees (Samet 1989) index a network of linear spatial objects over its geometric space. Another approach to spatial indexing of such networks is to represent such networks as a node list with an edge list associated with each node. The nodes can then be indexed with point quad-trees (Samet 1989). The merits of these spatial indexing schemes for access in main memory as well as access in secondary storage is worth investigating.

7 Summary

We have identified a broad issue within spatial database research namely handling spatially embedded graphs and presented some important problems that must be addressed in relation to it. Most of these are discussed in isolation in the literature. The author is working on the issue of subgraph maintenance as a tool for solving location problems.

References

ESRI International, editor (1992). Arc/Info Network Module, User Manual ESRI, 1992.

Guting, R H (1994). Grapdb: Modelling and querying graphs in databases. In 20th Int. Conf. on Very Large Databses, Santiago, 1994.

Newell, R G and M Doe (1995). Discrete geometry with

seamless topology in a GIS. Technical Report 23, Smallworld Systems Inc., 1995.

Mainguenaud (1995). Modelling the network component of geographical information systems. Int. J. Geographical Information Systems, 9(6):575-593, 1995.

Dueker, K J and R Vrana (1992). Dynamic segmentation revisited: a milepoint linear data model. URISA Journal, 4(2):94-105, 1992.

Naguleswaran, M and W Kreutzer (1996). Towards a domain framework for modelling linear spatial objects in urban systems. In Software Engineering Education & Practice, Otago, NZ, 1996.

Samet, H (1989). Design and Analysis of Spatial Data Structures. Adison-Wesley, 1989.

A comparison between land classification based on interpretations directly from aerial photographs and field mapping using a GPS system

M.Schlup[1,2], C.C. Boswell[2], G.K. Hutchinson[2]
[1] 22 Waldeggst., 2540 Grenchen, Switzerland
[2] AgResearch, Invermay Agricultural Centre, PB 50034, Mosgiel, New Zealand

Presented at the second annual conference of GeoComputation '97 & SIRC '97, University of Otago, New Zealand, 26-29 August 1997

Abstract

Two different methods of land classification are compared on the same area of land. The traditional method of classifying land from aerial photographs is compared with a ground-based classification using a Global Positioning System (GPS). Both methods are utilised as separate layers in a Geographical Information System (GIS) to produce a geocoded database of two tussock grassland block areas on Tara Hills High Country Research Station, Omarama, New Zealand. This was part of a larger study where productivity, utilisation by merino sheep, and botanical composition of ecological units within high country pasture are being measured.

Morphological structures and some vegetational patterns can easily be identified on aerial photographs but are not always representative of ecological units. Detailed ground-based surveys are often required to produce large scale land classification maps of smaller areas. GPS used in conjunction with ground-based surveys appears to be a powerful and very effective mapping tool, with additional benefits for geocoding; but is more expensive.

Introduction

Traditionally land use classification in New Zealand high country is broadly based on aerial photography. For example soil mapping to a scale of 1:253,440 (4 mile to 1 inch) and land use capability mapping was based on aerial photographs and associated ground-truthing.

Interest in accurate and quick identification of ecological units within fenced blocks of high country pasture for a better understanding of the requirements for, and risks to their sustainability, lead to a consideration of alternative mapping techniques incorporating ground based geocoding of boundaries.

A traditional map derived from aerial photographs was compared with a map produced by ground based field mapping using differentially corrected data captured with the Global Positioning System (GPS) for the same area of land.

GPS is a US government maintained network of 24 satellites which are constantly emitting signals for global reception. Through trilateration, ranging, accurate timing and with satellite position information, ground location of receivers can be determined to an accuracy of about +/- 10 m. GPS allows rapid creation and updating of GIS (Geographical Information Systems) databases now commonly used in resource management and other fields.

Three-Dimensional Computer-Based Gold Prospectivity Mapping using Conventional Geographic Information systems, Three-Dimensional Mine Visualization Software and Custom Built Spatial Analysis Tools

Eduardo Videla[1] and Carl M. Knox-Robinson[2]
Centre for Teaching and Research in Strategic Mineral Deposits
Department of Geology and Geophysics
The University of Western Australia
NEDLANDS WA 6907, Australia
evidela@geol.uwa.edu.au
carl@geol.uwa.edu.au

Presented at the second annual conference of GeoComputation '97 & SIRC '97, University of Otago, New Zealand, 26-29 August 1997

Effective mineral exploration requires a detailed knowledge of the factors and processes which result in the formation of economic deposits. To apply this knowledge, a sound three-dimensional understanding of the geology and structure of a region is required. In most cases, surface geology can be mapped with a high degree of accuracy, however, the geology at depth has to be inferred from geophysical methods, or through drilling programs, and is therefore mapped at much lower spatial resolution. This anisotropy in spatial-data quality, coupled with the scarcity of three dimensional geographic information systems (GIS), makes computer-based exploration at a camp- or district-scale very difficult.

The process of computer-based three-dimensional mineral exploration is being addressed at the Centre for Teaching and Research in Strategic Mineral Deposits within the Department of Geology and Geophysics at The University of Western Australia. A research project is in progress which attempts to define a three-dimensional gold prospectivity model for the Wiluna goldfield. The aim of this research is to gain a better understanding of the factors which spatially control the location of the known ore bodies, and especially of high-grade zones within these bodies. Also, the research aims to identify potential continuations of known ore bodies, and to attempt to locate new prospective areas for gold mineralisation further to the south of the present goldfield.

The Wiluna goldfield comprises a region approximately 3km x 5km, and is situated in the northern part of the Archaean Yilgarn Block of Western Australia, approximately four kilometres south of the Wiluna townsite. The area has been mined for gold since the early 1900s, and presently comprises 11 open-cut and underground gold mines. Geological information for the region includes detailed surface mapping and over 6,000 unevenly distributed drill holes, totalling in excess of 300km of core samples. The drill-core has been assayed for gold and gold-related elements, including arsenic, antinomy, and sulphur. Plans and sections from mine construction activities provide detailed three-dimensional information for limited areas. These data have been entered into a 3D mining package, and a model of the surface and interpolated sub-surface geology and structure constructed. This model will be used as a base on which to conduct a gold prospectivity analysis

Several GIS-based methods have been developed to as-

sess and map prospectivity on a global to district scale using datasets such as geological maps, aeromagnetic and radiometric data, topography, and satellite imagery. Prospectivity mapping methodologies can be split into two broad groups: knowledge driven and data driven. Knowledge driven methodologies involve the application of conceptual models to appropriate spatial datase[illegible] eas data-driven methodologies look for significant [illegible] a-tionships between known sites of mineralisat[illegible] [illegible] sur-rounding geological features. Identified spatial relationships are quantified as mappable criteria and are ultimately integrated into a single prospectivity map. Techniques applied with varying degrees of success include Boolean logic, index-overlay, Bayesian statistics, fuzzy logic and artificial neural networks.

The majority of assessments to date have been two-dimensional in nature and normally conducted at a scale where deposits can be adequately represented as point features. This research has the additional complications in that the third dimension must be addressed, and that the scale of observation is such that ore-bodies have a definite volume, and cannot be regarded as simple point objects. Although present mining packages are good at visualising three-dimensional bodies, and are capable of measuring lengths, areas and volumes, most packages lack an in-built macro language which would allow a quantitative examination of gold prospectivity. Consequently, dedicated data handling programs are being developed to extract the spatial information from the mining software and to conduct quantitative spatial analysis techniques to identify and quantify significant spatial relationships between high-grade ore zones and the surrounding geology.

1- Objectives

Through the integration of two-dimensional surface geological maps and-three-dimensional subsurface information of the Wiluna goldfield, the aim of this research is to construct a three-dimensional geological model of the area and to quantitatively analyse controls on gold mineralisation. Use of these controls to define methodology for regional GIS-based gold prospectivity analysis.

2- Data collection

Data available for this project includes:

a) a detailed geological surface map at scale 1:2 500 produced in a previous research (S. Hagemann, 1992). This map is available in digital format and was updated with more recent information. It includes main lithostratigraphic units, first-, second- and third-order structures, and measurements of azimuth and dip of faults.

b) an extensive drillhole database including exploration and evaluation records. The information contained in the database includes-
   - > 6000 drillholes (RC, diamond, evaluation)
   - > 375 000 meters of drill
   - > 200 000 Au assays
   - > 95 000 geochemical assays
   - > 95 000 magnetic susceptibility of host rock measurements
   - > 15 000 rock descriptions

detailed geological maps of pit and underground works and interpreted geological cross sections.

3- Data inte[illegible]d 3D modelling

To achieve the best [illegible] v[illegible]ualization of the complex environment of the geological subsurface, the available information was integrated using a mining visualisation software. A 3D model was construct[illegible] [illegible]rrelating the surface map with the drillhole information, u[illegible]dergr[illegible] [illegible]d mining maps and interpreted geological section[illegible].

A number of problems need to be addressed in th[illegible] proc ess of interpretation and integration of data. These problems are technological as well as inherent to the data. Present mining visualisation software require high specialisation and the process of updating the model according to new information is difficult and extremely time consuming. In terms of the data, good correlation is achieved in dense sampled areas but an increasing degree of interpolation and uncertainty is introduced in poorly sampled areas combined with the inherent anisotropy and complexity of the geologic subsurface.

4- Controls of the mineralisation

It is accepted that the gold mineralisation is late in the tectonic evolution of the Yilgarn craton (Groves, 1993). The fault system, along with relevant lithological contacts, is the principal control of the mineralisation. As the faults play a critical role in the siting of the ore bodies, an accurate spatial representation of this structures is required.

The traces of the faults on the surface map, their projection in the underground mining maps, and the drillhole intersection of the fault in subsurface, provide the lines and points used to create these entities in space. An empirical spatial resolution of 15 m was adopted for the basic cells that represent three dimensional geological solids. Lines and points were spatially gridded at this resolution, and a best polynomial algorithm fitted a TIN (triangulated irregular network) to these points. In all cases, control points were left without gridding to validate the interpolation accuracy. The final surfaces look smooth and realistic and serve as a basis for further spatial analysis.

5- Data extraction

In this study the kind of datasets required for analysis are dependent primarily on the type of deposit investigated. The Wiluna lode gold deposits are predominantly structurally controlled with relative lithologic control. Consequently, the solid 3D representation of faults and structures of first to third order is required to identify suitable relationships with gold mineralisation and specially of high-grade accumulations within these bodies.

From the final 3D model, the TIN representation of solid geologic entities can be exported in various formats for further analysis. For this project, ASCII files of the format {x1,y1,z1,x2,y2,z2,x3,y3,z3} which represent the spatial coordinates of the three corners of the basic triangular units, were extracted for each entity.

6- Spatial analysis of the fault system

Considering that the ore bodies are mainly controlled by faults and are present in determinated sites, but not in others, it is inferred that particular geometrical features along these faults are responsible for gold deposition along with fluid-wallrock interaction and physico-chemical conditions in the time of the mineralisation. Extensional veins, dilational jogs, shear veins, divergent bends, etc, are all terms related with the geometry of the structures formed after the application of a directed regional stress to the rockmass. A measure of this deformation is the displacement along a fault, the azimuth and dip, the angle formed between faults, veins, joints, the orientation of the schistosity, etc.

In order to find relationships between gold mineralisation and the hosting structures, it is necessary a spatial discretisation of the faults into basic components, at a scale relatively similar to that of the gold assays, and to generate new variables relating the relative spatial position of gold and structure.

By construction, the fault surface is made from a variable number of ordered triangular facets connected by the sides, they represent the topology of the physical surface. The computation of the centroid (properly called hypocenter) of every triangular facet generates the points necessaries for the analysis.

At the same time, operating on the normal vector to every facet it is possible to calculate its spatial orientation, in terms of azimuth and dip. Angular relations between facets or their normals, allow measures of coplanarity, concavity, convexity, bends and variability in azimuth and dip of the faults or lithological contacts. The vector equation of the plane for every single facet enables to discriminate points in space relative to this plane in terms of above and below, or in geological terms hangingwall and footwall.

The computation of the distance to the nearest facet in the fault for every gold value in space, generates a spatial variable that relates gold grade with azimuth, dip and proximity to the fault.

7- Customised software

To conduct quantitative spatial analysis to identify significant relationships between high-grade ore zones and the surrounding geology, dedicated data handling programs were developed. These specific pieces of software were created in Borland C++, to fulfill the necessity of spatial

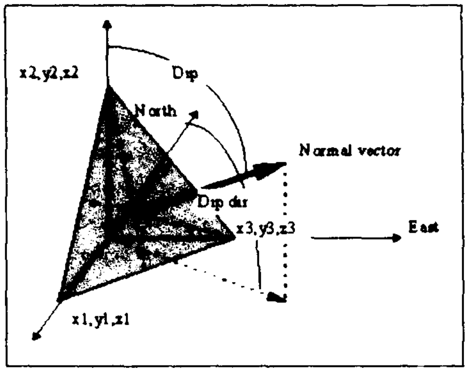


Figure 1 Spatial representation of a facet and its trigonometric attributes

analysis tools other than those which are standard in the mining software.

The name of each module, use and a sample of output chart is described below,

FACET-3D: for every facet on a fault surface, FACET-3D calculates the spatial coordinates of the centroid, azimuth, dip, dip direction, normal vector, director cosines, vectorial equation of the plane and fault identification (See Figure 1).

DIST-3D: for a set of spatially distributed gold assays, DIST-3D computes the shortest distance to a facet in the nearest fault. Discrimination between assays in the footwall and hangingwall is made through the sign of the relative distance. A (+) distance indicates points in the hangingwall and (-) in the footwall.

LAG-ASSAY: for a selected lag interval *h*, LAG-ASSAY computes the average and frequency of gold assays within this lag distance relative to the nearest fault position, increasing the searching distance away from the fault surface until all assays are exhausted. The averages are calculated for a normally distributed population of assays as well as a three-parameter lognormally distributed population. In this case the lag *h* controls the amount of smoothing of the distribution and hence is called smoothing factor (See Figure 2).

DIP-ASSAY: for a selected dip interval of facets in the fault surface, DIP-ASSAY computes the average and frequency of gold assays associated to these facets. The averages are calculated for a normally distributed population of assays as well as a three-parameter lognormally distributed population (See Figure 3).

AZIM_ASSAY: for a selected azimuth interval of facets in the fault surface, AZIM_ASSAY computes the average and frequency of gold assays associated to these facets. The averages are calculated for a normally distrib-

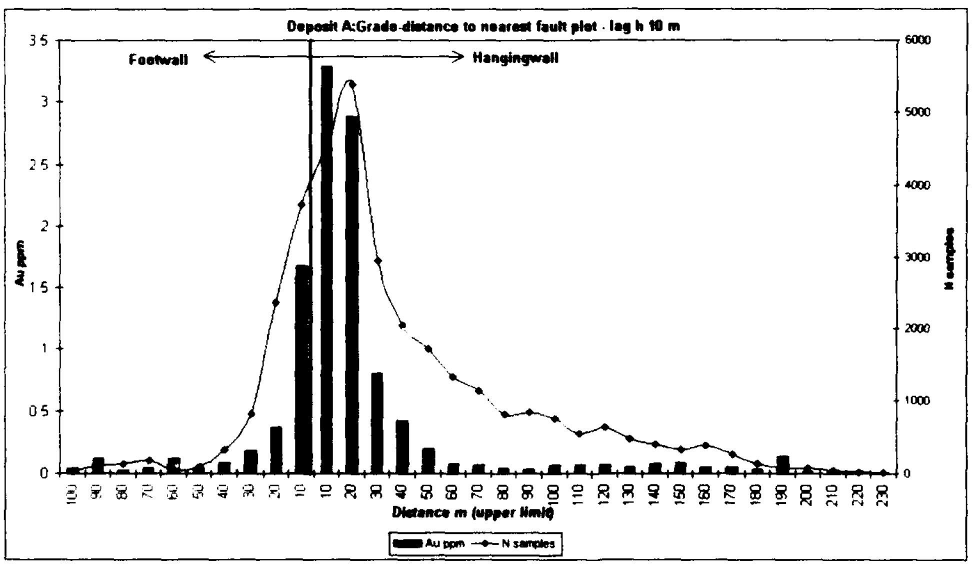


Fig 2 Lag-assay chart, proximity analysis

uted population of assays as well as a three-parameter lognormally distributed population (Figure 4).

DEPTH-ASSAY: for a selected depth interval, DEPTH-ASSAY computes the average and frequency of gold assays within this interval relative to the surface level. The averages are calculated for a normally distributed population of assays as well as a three-parameter lognormally distributed population (Figure 5).

ROCK-ASSAY: for every rock-type present at the mineralised site, ROCK-ASSAY computes the average and frequency of gold assays within this rock. The averages are calculated for a normally distributed population of assays as well as a three-parameter lognormally distributed population (Figure 6).

STRIKE-BIN: for a selected portion of a fault, STRIKE-BIN splits and bins the assays at selected distances from an origin and computes the average and frequency of gold assays within these bins designed perpendicular to the fault strike. The averages are calculated for a normally distributed population of assays as well as a three-parameter lognormally distributed population (Figure 7).

DIP-AZIM: for selected intervals in dip and azimuth of facets in the fault, DIP-AZIM computes the frequency of facets within these intervals for further statistics.

8- Preliminary results

One important mine "Deposit A", is examined using these techniques to quantify spatial relationships between gold mineralisation and structural features.

8.1- Proximity relationships

For Deposit A, a proximity relationship is identified between high-grade gold mineralisation and the portion of the fault hosting that mineralisation. Figure 2 shows that at a smoothing factor (*h*) of 10 meters, gold is concentrated in economic grades in a narrow corridor around the hosting fault. The distribution is asymmetric with the highest grades in the hangingwall up to 20 m away from the fault surface. In contrast the mineralisation in the footwall is less intense and restricted to the first 10 m., although the sampling frequency is less abundant in this portion of the fault.

For the discovery of parallel or secondary mineralised structures relatives to the main fault, the factor *h* has to be related to the sample size *n* and to the dispersion of the data. The more data available, the more precise is the search for details of the underlying density function.

Figure 8 demonstrates the effect of the factor (*h*) on a

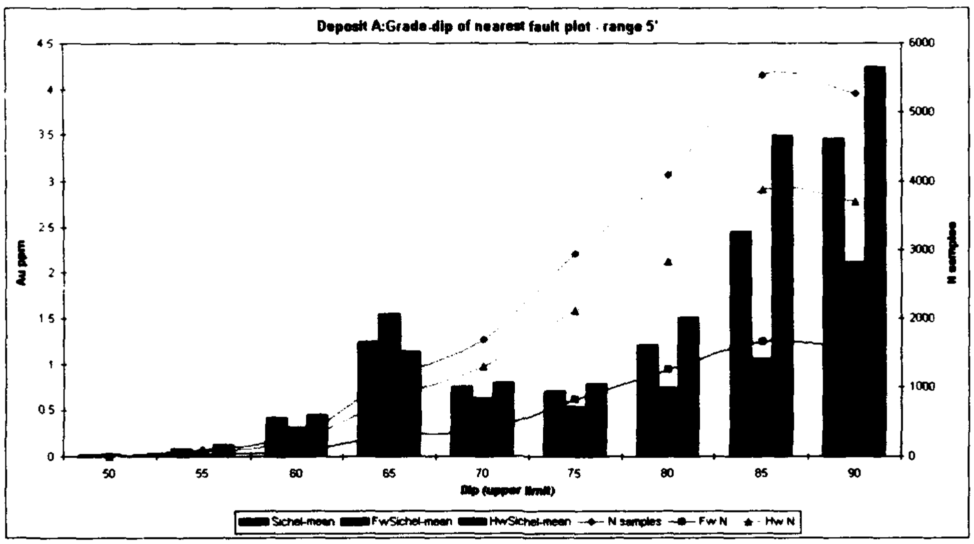


Fig 3. Dip-assay chart

density estimate of gold-distance to fault. (h = 2.5 m). From Figure 1 it is known that the first 20 m away from the fault, in the hangingwall, is highly mineralised, using h = 2.5 m it is possible to detect two discrete zones between 5 and 7.5 m and 12.5 and 15 m away from the fault accounting for most of the gold in the first 20 m. These peaks correlate in depth with two parallel structures hosting high grade mineralisation in the south portion of the deposit. These minor structures were not incorporated in the model, but are highlighted using the appropriate h. See fig 7, bins 150

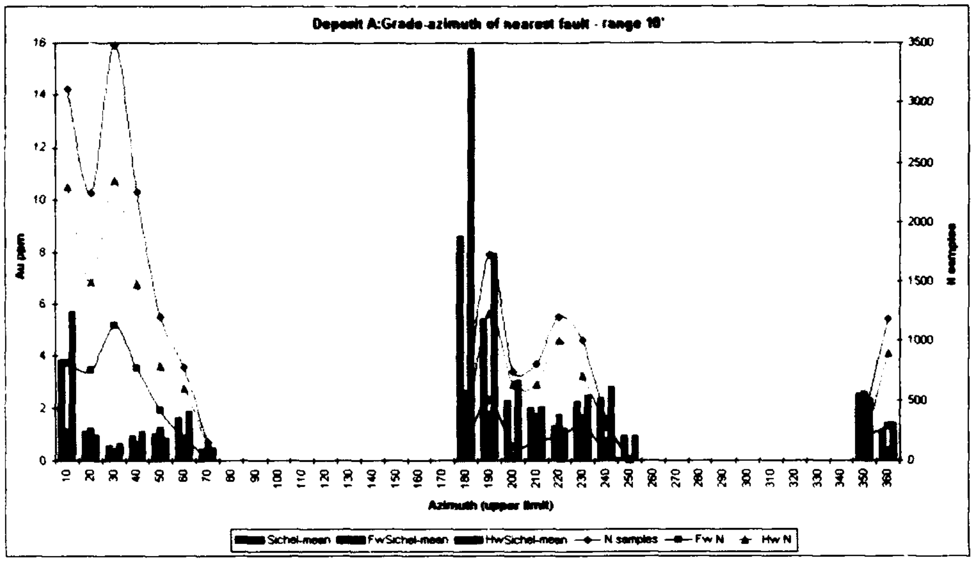


Fig 4. Azimuth-assay chart

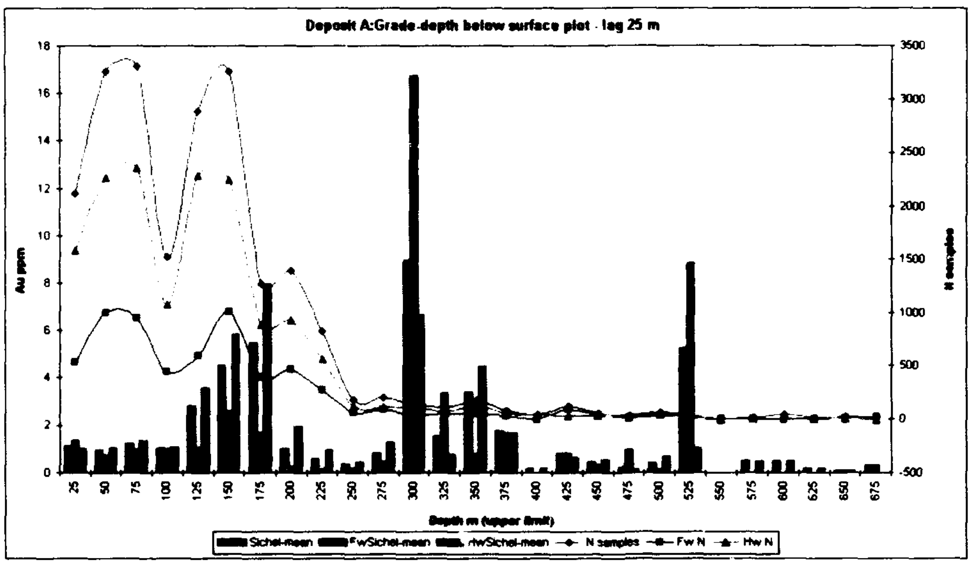


Fig 5. Depth-assay chart

and 200 for the along-strike extension of this features.

A buffer zone both sides of the fault can be distinguished according to appropriate cut-off grades to limit the lateral extension of the mineralised body.

8.2- Dip-Azimuth relationship

Identified a proximity relationship between gold mineralisation and the hosting fault, sections of the fault striking and dipping in a restricted interval of directions may be more prospective than others.

Applying an appropriate cut-off for high-grade assays, an *i* x *j* contingency table of frequency of facets within a par-

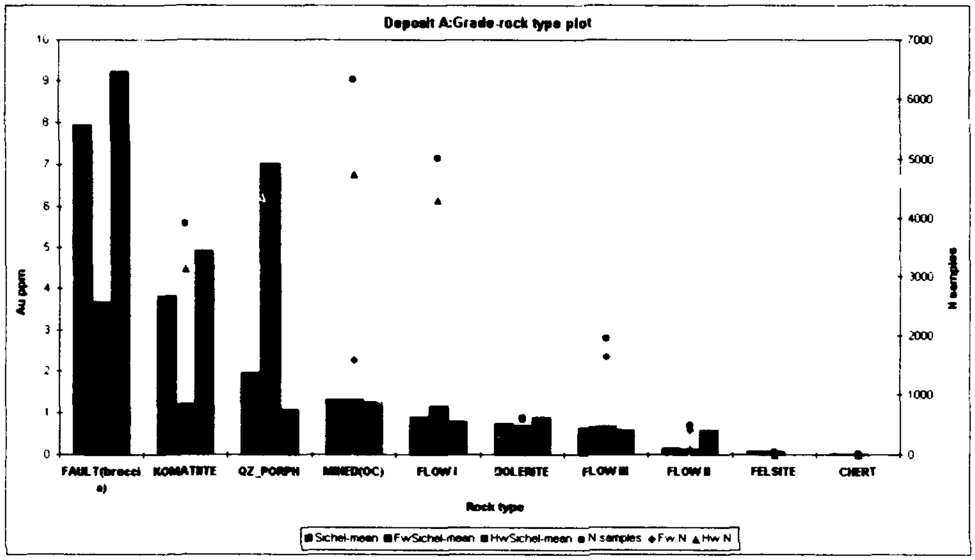


Fig 6 Rock-assay chart

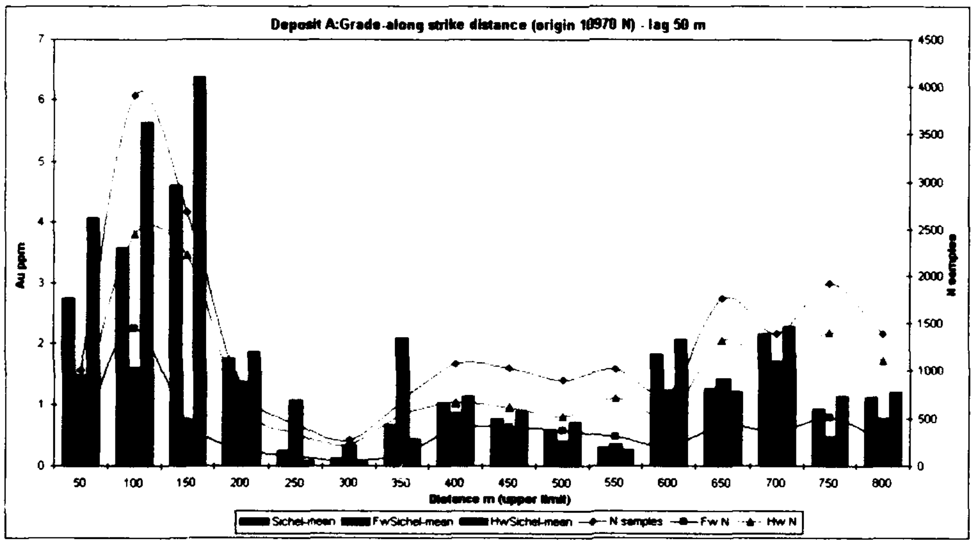


Fig 7 Strike-bin chart

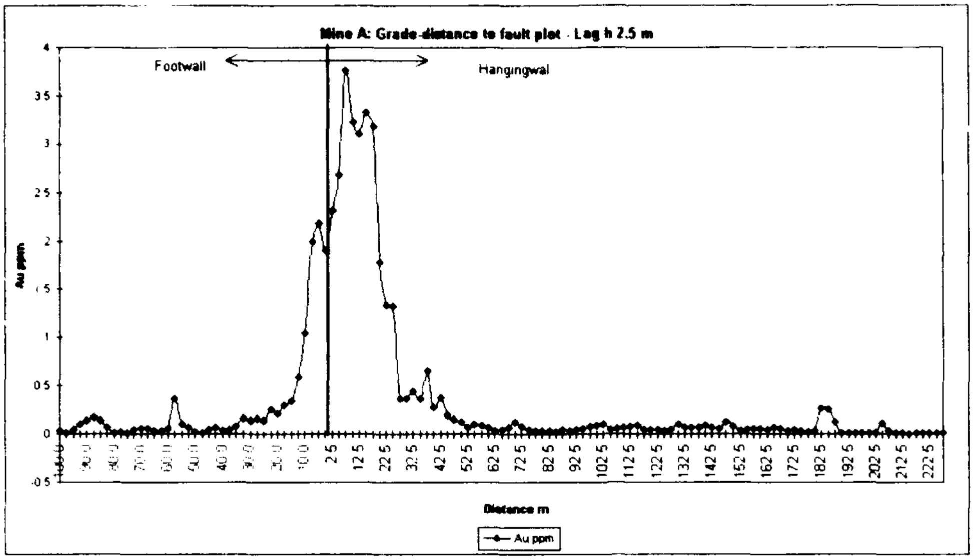


Fig 8. Effect of choice of smoothing factor h on a density estimate of gold-distance to fault. (h = 2.5 m)

ticular interval of azimuth and dip can be constructed. Based on the relative proportion of facets in the fault, the expected number of facet related to high-grade can be calculated. These values are the expected if the position of the high-grades assays is independent of strike and dip.

For statistical reasons the expected value E_{ij} for each interval should be greater than 1.0 without endangering the validity of the test (Conover, 1980). The cells in the contingency table of expected values with frequencies less than 1.0 are condensed into a fewer number of contiguous and logically arranged cells, so that no cell contains less than 1.0 expected facet. The same arrange of cells is then applied to the observed O_{ij} contingency table. The observed and expected tables can then be compared using a Chi-square test for independence with m-1 degrees of freedom, being m the number of condensed cells.

The test statistic c^2 is given by

$$\chi^2 = \Sigma\, \Sigma \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$$

$$\text{where } E_{ij} = \frac{n_i C_i}{N_i}$$

n_i represents the number of facets for each interval of azimuth-dip in the fault, and C_i/N_i the proportion of high-grade related to total facets for the deposit.

If a Dip-Azimuth to high-grade gold relationship is found to exist, the Chi-squared component of each Dip-Azimuth category can be examined to determine which particular combination of Dip-Azimuth is more prospective for high-grade values.

An example of contingency tables for a Chi-square test for independence between observed and expected facets of particular azimuth and dip associated to gold assays > 5 ppm, and critical values is shown in Figure 9.

As the statistic is larger than the critical Chi-square value, the null hypothesis that both distributions are identical is rejected at a confidence level of 95%, and consequently a high-grade dip-azimuth relationship is established.

Circular or spherical statistical analysis in case of azimuth or azimuth and dip data, is required to assess the deviation about the mean direction vector of the fault, of these more prospective sites. These departures can then be used as a predictive tool in the search for extensions of present deposits or to target new ones in similar geological conditions.

9- Conclusions

Quantification of controls on gold mineralisation at camp-scale requires a sound three-dimensional understanding of the geology and structures involved. Integration of detailed surface geological maps with subsurface underground mining and direct drilling information, lead to the construction of acceptable representations of the three-dimensional geology of the area. These models are used as a base on which to conduct gold prospectivity analysis. Dedicated data handling programs are designed to quantify and analyse spatial relationships that control known ore bodies. Characteristic features can be identified as more prospective and consequently used as a predictive tool in the location of new deposits.

References

Conover W.J. (1980) Practical non-parametric statistics, 2ed. John Wiley & Sons, pp 153-168.

Groves D.I. (1993) The crustal continuum model for late-Archeaen lode-gold deposits of the Yilgarn Block, Western Australia, Mineralium Deposita 28. pp 366-374.

Houlding S.W. (1994) 3D Geoscience Modeling. Computer Techniques for geological characterization. Springer-Verlag, New York.

Knox-Robinson C.M. (1994) Archeaen lode-gold mineralisation potential of portions of the Yilgarn Block, Western Australia: development and implementation of methodology for the creation of regional-scale prospectivity maps using conventional geological map data and a geographic information system (GIS). Unpublished PhD Thesis, The University of Western Australia.

Mardia K.V. (1972) Statistics of directional data, Academic Press, London, UK 1972.

Rock N.M.S. (1988) Numerical Geology: a source guide, glossary, and selected bibliography to geological use of computers and statistics. Lecture Notes in Earth Sciences 18, Springer-Verlag, New York, 427pp.

Statistics

Dip	50	55	60	65	70	75	80	85	90	N
Az										
10	1 398924		1 614143	0 969831	0 08553	1 320394	0 063667	3 39528	**8.2647**	17
20			0 002051	0 132209	4 192078	**7.048424**	1 144453	1 786373	0 0001	14
30				0 010762	0 307329	0 005041	0 822667	1 570561	0 709	3
40			1 990776	0 033822	0 371655	0 036314	0 120604	3 606979	**7.5084**	14
50	1 452729			**6.700588**	2 262688	5 1903	2 785874	0 307269	0 0292	19
60	0 414187			**37.27235**		**11.43937**	**14.61466**	**13.90317**	**21.821**	99
70										
80				1 614143						2
180										
190								**53.66632**	**31.347**	85
200						0 029868		0 004934	0 4955	1
210				0 042509		4 788624	1 479279	0 007178	0 0096	6
220						0 100272	1 972372	1 995913	5 2634	9
230				1 932043		1 344955	0 491707	0 003888	**10.415**	14
240						1 291314		1 647967	**10.651**	14
250										
260										
270										
340										
350						0 935073				1
360				0 014606		0 789812	0 118598	0 071892	0 0215	1
	3		4	49	7	34	24	82	97	**299**

N. cells=	**70**	Critic value for individual cells(0 95)=	5 99
D.F.	**69**		
δ (0.95)=	**90**		

Fig. 9 Contingency table for Chi-square test of independence